EVOLUTIONARY IMPACTS OF DNA METHYLATION ON VERTEBRATE GENOMES

A Dissertation
Presented to
The Academic Faculty

by

Navin Elango

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Bioinformatics in the
School of Biology

Georgia Institute of Technology
December 2008

EVOLUTIONARY IMPACTS OF DNA METHYLATION ON VERTEBRATE GENOMES

Approved by:

Dr. Soojin Yi, Advisor
School of Biology
Georgia Institute of Technology

Dr. John McDonald
School of Biology
Georgia Institute of Technology

Dr. Michael Goodisman
School of Biology
Georgia Institute of Technology

Dr. Eric Vigoda
College of Computing
Georgia Institute of Technology

Dr. Kirill Lobachev
School of Biology
Georgia Institute of Technology

Dr. James Thomas
Department of Human Genetics
Emory University

Date Approved: August 7, 2008

To my dear parents Elango and Sumathi…

ACKNOWLEDGEMENTS

I am deeply grateful to my parents for their boundless patience and support while I followed my desired path to obtain a PhD. I sincerely thank my advisor Dr. Soojin Yi for her patience, infallible support and friendly encouragement, which made this path truly enjoyable.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BAC	Bacterial Artificial Chromosome
BLAST	Basic Local Alignment Search Tool
BLAT	Blast Like Alignment Tool
bps	Basepairs
CpG	Cytosine immediately followed by Guanine in 5' to 3' direction
CGI	CpG Island
DNA	Deoxyribonucleic Acid
DNMT	DNA Methyl Transferase
EMBL	The European Molecular Biology Laboratory
EST	Expressed Sequence Tag
GO	Gene Ontology
GpC	Guanine immediately followed by Cytosine in 5' to 3' direction
HCG	High CpG Content
LCG	Low CpG content
LCGI	Long CpG Island
LINE	Long Interspersed Nuclear Element
LTR	Long Terminal Repeat
Mbps	Mega base pair
MYA	Million Years Ago
NCBI	National Center for Biotechnology Information

NCGI	No CpG Island
Polr2a	RNA Polymerase II
RefSeq	Reference Sequence database
RNA	Ribonucleic Acid
SCGI	Short CpG Island
SINE	Short Interspersed Nuclear Element
TE	Transposable Element
TSS	Transcription Start Site
UCSC	University of California Santa Cruz

SUMMARY

DNA methylation is an epigenetic modification in which a methyl group is covalently added to the DNA. In vertebrate genomes methylation occurs almost exclusively at cytosines immediately followed by a guanine (CpG dinucleotides). Two important aspects of DNA methylation have inspired several recent scientific investigations including those in this dissertation. First, methylated cytosines are hotspots of point mutation due to a methylation-dependent mutation mechanism, which has caused a deficiency of CpGs in vertebrate genomes. Second, DNA methylation in promoters is linked with transcriptional silencing of the associated genes.

This dissertation presents the results of four studies in which I investigated the evolutionary impacts of DNA methylation on vertebrate genomes. The first two studies investigated the impacts of DNA methylation on neutral evolution. The third and fourth studies investigated the functional impact of DNA methylation on the evolution of vertebrate promoters. The research advances achieved through my studies are presented below.

Research advance 1: Understanding the rate of neutral substitutions in genomes has been an area of long-standing interest among evolutionary biologists. Two disparate views exist among scientists in regard to the rates of neutral single nucleotide substitution (molecular clock) across evolutionary lineages: 1) rates of single nucleotide substitution are approximately equal across lineages (the time-dependent *molecular clock hypothesis*), and 2) rates of single nucleotide substitution vary among lineages in a generation-time

dependent manner (*the generation-time effect hypothesis*). The first study of my dissertation provides evidence that primate genomes exhibit both time-dependent and generation-time dependent molecular clocks depending on the region of the genome considered. In particular, the results demonstrate that methylation dependent substitutions at CpG sites exhibit a time-dependent molecular clock, while substitutions at non-CpG sites exhibit a generation time dependent molecular clock. A simple mathematical model was developed to study the effect CpG substitutions on the molecular clock, and the predictions of the model are discussed with respect to the two disparate views that prevail among scientists.

Research advance 2: The second study investigated the importance of methylation dependent CpG mutations in intra-genomic substitution rate variation. The results of this study show that the rate of substitution at CpG dinucleotides is negatively correlated with local G+C content of the region. The strength of this correlation decays with increasing distance from the CpG site, and extends up to ~ 1500 – 2000 bps on each side of the site. This pattern is not observed in the case of mutations at non-CpG sites. These observations are attributed to a local DNA strand separation required for methylation-dependent CpG mutations to occur, and discussed in context of the importance molecular origins of mutations in genome evolution.

Research advance 3: Human promoters fall into two structural classes based on their CpG content – the heavily methylated low-CpG (LCG) class and the hypo-methylated high-CpG (HCG) class. HCG promoters are associated with genes expressed more broadly compared to genes with LCG promoters. The third study investigated the impact of DNA methylation on the evolution of vertebrate promoters by analyzing

genomes of several chordates including sea squirt, zebrafish, frog, chicken, and human. The results show that the two structural classes of promoters are a common feature in vertebrate genomes but not in the invertebrate outgroup (sea squirt). The association of HCG class with broadly expressed genes and LCG class with tissue-specific genes is also conserved among vertebrates. A model for the evolution of the two classes of vertebrate promoters is proposed.

Research advance 4: Vertebrate genomes are heavily methylated, with a large fraction of CpGs exhibiting methylated cytosines, except for certain regions called CpG islands (CGIs) that generally remain unmethylated. Because of its unmethylated state CGIs are CpG rich and often colocalize with HCG promoters, and are typically thought of as markers for broadly expressed genes. The final study of my thesis provides evidence that not all promoter associated CGIs are made equal. In particular, the results show that long-CGI promoters are associated with genes that exhibit intermediate levels of tissue specificity. The tissue specificity of genes with long-CGI promoter lies between those with short-CGI promoter (widely expressed) and those with no-CGI promoter (highly tissue specific). Long-CGI promoters are enriched with highly conserved genes involved in important biological processes like development, transcription regulation, and signaling. Long-CGI promoters also contain a larger number of RNA polymerase II binding sites, which are occupied in an intermediately tissue specific manner. These observations suggest that long-CGI promoters are associated with genes that require more complex regulation strategies.

CHAPTER 1

INTRODUCTION AND MOTIVATION

INTRODUCTION TO DNA METHYLATION AND POINT MUTATIONS

Understanding the mechanisms, rates, and patterns of mutations-the ultimate source of genetic variation within and between populations-is quintessential to comprehend the phenomenon of evolution. The most common type of mutation that occurs in nature is the *point mutation*, in which a single nucleotide in the DNA is incorrectly replaced by another nucleotide. Because of its sheer abundance, single nucleotide mutations play critical roles in genome evolution.

The past few decades have witnessed growing interest in an epigenetic mark called DNA methylation, particularly because of its profound influences on genome organization, transcription regulation, and ability to induce single nucleotide mutations (COSTELLO and PLASS 2001; KLOSE and BIRD 2006; SUZUKI and BIRD 2008; SUZUKI *et al.* 2007). This dissertation investigates some evolutionary consequences of DNA methylation-dependent single nucleotide mutations on vertebrate genomes. This chapter provides a brief introduction to DNA methylation, and evolutionary aspects of point mutations. In accord with the objective of this dissertation much of the introduction is focused on the vertebrate lineage, especially on mammals (rodents and primates) where these topics have been studied extensively. Details from other taxa are provided, where necessary, to stress the similarities, variations, and importance of DNA methylation pattern in other forms of life.

DNA methylation – an epigenetic mark

DNA methylation is an epigenetic modification in which a methyl group is covalently added to the DNA. It is found in genomes of diverse organisms including prokaryotes and eukaryotes. In prokaryotes DNA methylation occurs on both cytosine and adenine bases, whereas in multicellular eukaryotes it appears to be confined to cytosine bases [Table 1.1; (SUZUKI and BIRD 2008)]. An additional restriction exists in vertebrate genomes, where DNA methylation is confined to cytosines immediately followed by a guanine (called as CpG dinucleotides or CpG sites).

DNA methylation patterns differ across various taxa

The pattern of DNA methylation varies significantly between taxa (Table 1.1). In vertebrates, DNA methylation is found throughout the genome (~80% of the CpG dinucleotides methylated) except for short regions called *CpG islands,* which remain unmethylated in several tissues. CpG islands encompass a minor fraction of vertebrate genomes. In mammals, for example, these unmethylated regions encompass a mere 1-2% of the genome (BIRD *et al.* 1985; BIRD 1986; ILLINGWORTH *et al.* 2008). This *global* DNA methylation pattern is observed in several distantly related vertebrate genomes (e.g., *Homo sapiens*, *Gallus gallus*, *Xenopus tropicalis*, and *Danio rerio*), suggesting that it is the most frequent pattern among vertebrates (TWEEDIE *et al.* 1997).

Notably, the global DNA methylation pattern found in vertebrates is not ubiquitous among eukaryotes. The genomes of several organisms like the fungi *Saccharomyces cervisiae* and the worm *Caenorhabditis elegans* lack genes encoding proteins that makeup the DNA methylation machinery, and hence are devoid of

Table 1.1 DNA methylation pattern in multicellular eukaryotes

Species	Methylated Sites	Overall Pattern	Gene Body Methylation	Transposon Methylation
Plants				
Arabidopsis thaliana	CG, CNG and CNN	Mosaic	Yes	Yes
Zea mays	CG, CNG and CHH	Mosaic	N/A*	Yes
Oryza sativa	CG, CNG and CHH	Mosaic	Yes	Yes
Fungi				
Saccharomyces cerevisiae	None	None	No	No
Schizosaccharomyces pombe	None	None	No	No
Neurospora crassa	CNN	Mosaic	No	Yes
Ascobolus immersus	CNN	Mosaic	N/A	Yes
Insects				
Drosophila melanogaster	CT and CA	Sparse	Yes	Yes
Apis mellifera	CG	Mosaic	Yes	No
Myzus persicae	CG	Mosaic	Yes	N/A
Deuterostomes				
Echinus esculentus	CG	Mosaic	N/A	Yes
Strongylocentrotus purpuratus	CG	Mosaic	Yes	N/A
Ciona intestinalis	CG	Mosaic	Yes	Yes
Vertebrates				
Danio rerio	CG	Global	Yes	Yes
Xenopus laevis	CG	Global	Yes	Yes
Homo sapiens	CG	Global	Yes	Yes

* N/A: Data Not Available

methylation (SUZUKI and BIRD 2008). The insect *Drosophila melanogaster* exhibits a sparse DNA methylation pattern. For example, only 0.4% of the cytosines are methylated in *Drosophila melanogaster* embryos (SUZUKI and BIRD 2008).

The most common methylation pattern observed in the eukaryotes studied thus far is the *mosaic* DNA methylation pattern with long tracts of methylated DNA interspersed between long tracts of unmethylated DNA [Table 1.1; (SUZUKI *et al.* 2007; TWEEDIE *et al.* 1997)]. This pattern is observed in plants (e.g., *Arabidopsis thaliana, Zea mays*), fungi (e.g., *Neurospora crassa, Ascobolus immersus*), insects (*e.g., Apis mellifera, Myzus persicae*), and other invertebrate animals closely related to vertebrates *(e.g., Ciona intestinalis, Strongylocentrotus purpuratus*). Based upon these observations, it is proposed that the switch from mosaic to global DNA methylation pattern occurred during early vertebrate evolution (TWEEDIE *et al.* 1997).

Establishment of the DNA methylation pattern

DNA methylation is realized through enzymes called as DNA methyltransferases (DNMTs). Several prokaryotic and eukaryotic DNMTs have been characterized thus far (ADAMS 1995; GROMOVA and KHOROSHAEV 2003). As expected, many species that lack DNA methylation (e.g. *Saccharomyces cervisiae* and *Caenorhabditis elegans*) also lack functional DNA methyltransferases in their genome.

In the well-studied mammalian system four DNMTs have been identified *(DNMT1, DNMT2, DNMT3a, DNMT3b)*, which fall into two classes based on the preferred DNA substrate (KLOSE and BIRD 2006). The *maintenance methyltransferase DNMT1* prefers hemi-methylated substrate, and is responsible for copying pre-existing

methylation patterns to the new strand during DNA replication. On the other hand, *de novo methyltransferases* (*DNMT3a and DNMT3b*) are responsible for cytosine methylation at previously unmethylated CpG sites. *DNMT2* exhibits a weak methyltransferase activity, suggesting that it is not a major player in setting up DNA methylation patterns (HERMANN *et al.* 2003; OKANO *et al.* 1998).

Mammals also contain methyl-CpG-binding proteins that play crucial roles in decoding the epigenetic information encoded by DNA methylation. Six methyl-CpG-binding proteins have been characterized so far [Kaiso, MBD1, MBD2, MBD3, MBD4, MeCP2 (KLOSE and BIRD 2006)].

The establishment of DNA methylation pattern in mammals proceeds through several defined stages during development. Methylation levels in female germ cells are typically lower than that of male germ cells. The gamete methylation pattern is erased by a global genome-wide demethylation event that occurs close to the eight-cell stage of blastocyst formation (KAFRI *et al.* 1992; MONK *et al.* 1987). Next, a global *de nevo* methylation event that occurs during the implantation stage re-establishes the methylation pattern (KAFRI *et al.* 1992; MONK *et al.* 1987). The methylation pattern thus established is dynamic. In adults, DNA methylation pattern is known to vary significantly among different tissue types (ECKHARDT *et al.* 2006; OAKES *et al.* 2007).

Functions of DNA methylation

DNA methylation plays critical roles in diverse biological processes, the identity of which varies tremendously across taxa. In prokaryotes DNA methylation is a component of the restriction system, which protects the host from the effects of foreign

DNA (NOYER-WEIDNER and TRAUTNER 1993). Restriction endonucleases differentiate between host and foreign DNA by means of their DNA methylation pattern. Foreign DNA, which is void of DNA methylation is identified and cleaved by restriction endonucleases. DNA methylation is also involved in the control of replication fidelity in prokaryotes. Immediately after replication, the template strand is methylated whereas the newly synthesized strand is not methylated. When a mismatch is detected between these strands, the repair system takes advantage of this difference in methylation pattern to identify the newly synthesized strand and correct the mismatch (COOPER *et al.* 1993).

The role of DNA methylation in host defense extends to some eukaryotes (YODER *et al.* 1997). Several lines of evidence indicate that DNA methylation plays a crucial role in suppressing transposable elements. For example, a recent study showed that the methylated component of the *Neurospora crassa* genome consists almost exclusively of relics of transposons that were subject to repeat-induced point mutation- a genome defense system that mutates duplicated sequences [Table 1.1; (SELKER *et al.* 2003)]. Indirect evidence through sequence analyses suggests that transposable elements are heavily methylated in primate germline (MEUNIER *et al.* 2005). It has also been shown that DNA methylation is involved in silencing transgenes and viral sequences in mammals (COLLICK *et al.* 1988; KISSELJOVA *et al.* 1998; SASAKI *et al.* 1993).

DNA methylation is crucial for embryonic and post-birth development in mammals. Mouse embryos having homozygous deletion of DNA methyltransferases *DNMT1* and *DNMT3b* die before birth (LI *et al.* 1992; OKANO *et al.* 1999). Knocking out *DNMT3a* leads to death in ~ 4 weeks(OKANO *et al.* 1999). DNA methylation is the cause of several disease causing germline mutations (COOPER and YOUSSOUFIAN 1988), and

somatic mutations that lead to cancer (RIDEOUT *et al.* 1990). Aberrant promoter methylation has been shown to be associated with cancer (JONES and LAIRD 1999; OSHIRO *et al.* 2005). Methylation has also been implicated in immunodeficiency, centromeric instability, facial anomalies (ICF) syndrome. Recent studies have shown that the ICF patients have mutations in the gene encoding the DNA methyltransferase *DNMT3b*, which leads to specific chromosomal decondensation and reduced levels of methylation in satellite DNA(OKANO *et al.* 1999; XU *et al.* 1999). Rett syndrome patients have mutations in the methyl-CpG binding protein *MeCP2*, suggesting a possible role of the inability to decode methylation signals in causing this disease (AMIR *et al.* 1999).

Role of DNA methylation in mammalian gene regulation

In addition to the roles indicated above, DNA methylation is known to play important roles in mammalian gene regulation (KLOSE and BIRD 2006). Several lines of evidence suggest that DNA methylation is associated with gene silencing (BOYES and BIRD 1991; COMPERE and PALMITER 1981; KASS *et al.* 1997; SIEGFRIED *et al.* 1999). Many models have been proposed to elucidate the causal mechanism of this association (KLOSE and BIRD 2006). First, methylated cytosines in promoters of genes may directly interfere with the binding of transcription activators. Second, methylated cytosines may recruit methyl-CpG binding proteins, which can in turn recruit co-repressors to suppress the gene. Third, DNMTs are physically linked to histone modification enzymes. This suggests that DNA methylation may be coupled with histone modifications that confer closed chromatin structure, thereby rendering the associated gene inactive. Fourth, DNA methylation in gene body may interfere with transcription elongation (see (KLOSE and BIRD 2006) references therein).

The importance of DNA methylation in mammalian gene regulation and promoter evolution is highlighted by the fact that the promoters of protein coding genes in the human genome fall into two distinct classes, hypo-methylated and hyper-methylated, in several tissues and cell types including germline (SAXONOV *et al.* 2006; WEBER *et al.* 2007). The hyper-methylated promoters typically exhibit a lower CpG content compared to the hypo-methylated promoters due to methylation dependent hypermutability (see below). Therefore the hyper-methylated and hypo-methylated classes of promoters are also termed low-CpG (LCG) and high-CpG (HCG) classes, respectively (SAXONOV *et al.* 2006). An analysis of RNA polymerase II occupancy revealed that 66% of the hypo-methylated promoters were active, while the same measure for hyper-methylated promoters is 11% (WEBER *et al.* 2007). The hypo-methylated promoters typically contain CpG islands, and are associated with broadly expressed genes. The hyper-methylated promoters are associated with tissue-specific genes (SAXONOV *et al.* 2006; WEBER *et al.* 2007).

DNA methylation is also associated with dosage compensation in mammals. CpG islands in the promoters of several genes including house keeping genes like *HPRT*, *G6PD*, and *PGK1* are methylated in the inactive copy of the X-chromosome in mouse cells (KASS *et al.* 1997), presumably to achieve similar levels of expression in males and females. For many of these genes DNA methylation is preceded by silencing, suggesting that DNA methylation may play a role in the maintenance of silencing rather than initiating the event (COSTELLO and PLASS 2001).

Methylation dependent hypermutability of CpG dinucleotides

It is generally accepted that cytosine and 5-methylcytosine are vulnerable to spontaneous deamination (FRYXELL and MOON 2005; FRYXELL and ZUCKERKANDL 2000). Deamination of cytosine produces uracil (C -> U transition), which is readily identified and corrected by mismatch repair enzymes. However, deamination of 5-methylcytosine produces thymine (C-> T transition; Figure 1.1). The repair enzymes are less efficient in correcting C->T transitions, which in turn leads to the high rate of mutation at methylated cytosines (COULONDRE *et al.* 1978; EHRLICH *et al.* 1986; RAZIN and RIGGS 1980).

Apart from their dependence on DNA methylation, CpG mutations differ from mutations at other sites in two important aspects. First, a local DNA strand separation (DNA melting) is a pre-requisite for the deamination reaction. Therefore, the rate of CpG mutation is dependent on the local G+C content (FRYXELL and MOON 2005). CpG mutation rate is higher in low G+C content regions (which melt more readily) compared to that in high G+C content regions. Second, while most mutations in other sites seem to occur because of errors in DNA replication, CpG mutations occur spontaneously in germline. Therefore we hypothesize that it is independent of DNA replication, and provide evidence for this hypothesis in dissertation.

CpG mutations play crucial roles in genome evolution. The rate of methylation dependent mutations is ~10 – 50 fold higher than that of other single nucleotide mutations. In organisms where methylation is targeted to CpG dinucleotides (e.g. vertebrates), this leads to a dearth of CpG dinucleotides (FRYXELL and MOON 2005; SVED and BIRD 1990). The human genome, for example, contains only 22% of the CpG dinucleotides expected based on its G+C content. The importance of DNA methylation in

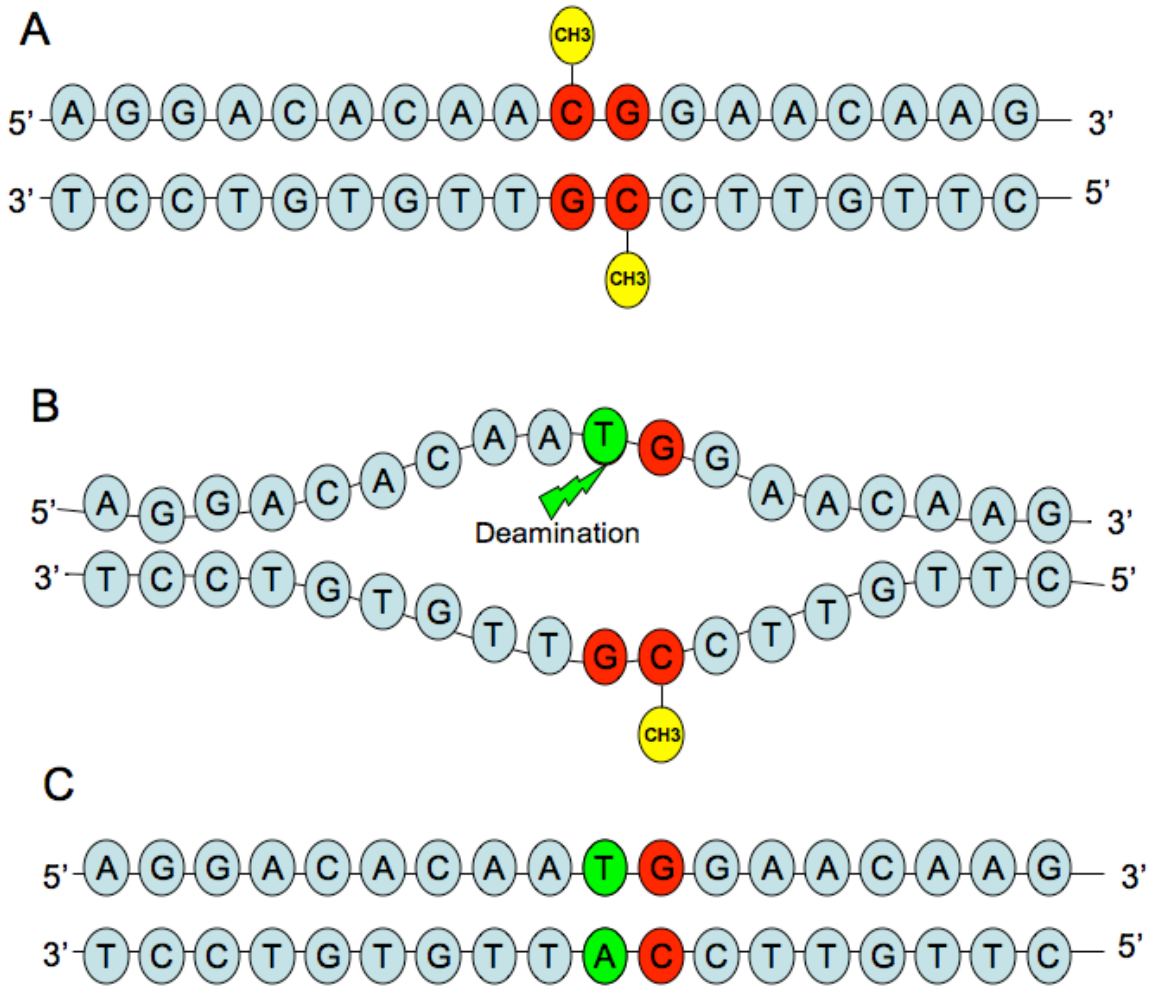


Figure 1.1. Hypermutability of methylated CpG dinucleotides. Methylated CpG dinucleotides (depicted as red circles in **A**) are highly vulnerable to deamination. Deamination requires a local DNA strand separation and mutates the cytosine to thymine (**B**). If left unrepaired, this may be passed on as a C to T mutation (**C**) to the next generation.

evolution is also demonstrated by the fact that ~25% of all single nucleotide substitutions between the human and chimpanzee genomes have occurred at CpG dinucleotides (MIKKELSEN T 2005). CpG mutations are also implicated in the evolution of mammalian isochores (FRYXELL and ZUCKERKANDL 2000).

Single nucleotide mutations and the molecular clock

In this dissertation I have investigated the role of DNA methylation-dependent point mutations on vertebrate genome evolution. However, investigations on the importance of single nucleotide substitutions in evolution began much before the significance of DNA methylation dependent mutations was realized. With the advent of rapid DNA sequencing methods in the early 1970's, comparative DNA sequence analyses started playing a central role in molecular systematics. The numbers of single nucleotide substitutions observed between homologous DNA sequences from different species were used as measures of their evolutionary distances to construct phylogenetic trees. Playing a pivotal role in such analyses is *the Molecular clock hypothesis*, which states that the rate of substitution (number of substitutions per year) does not vary across evolutionary lineages.

The molecular clock hypothesis was first formulated by Emile Zuckerkandl and Linus Pauling using protein sequences. They observed that the number of amino acid substitutions in hemoglobin between lineages scale approximately to divergence times estimated from fossil records, suggesting the existence of a molecular clock. They also noted that for a molecular clock to exist the amino acid changes must be limited to neutral or nearly neutral changes, because Darwinian selection may lead to rate variation across species. Several recent studies have also indicated the existence of a molecular

clock. For example (KUMAR and SUBRAMANIAN 2002) analyzed synonymous substitution rates in protein coding regions of several mammalian lineages and concluded that the rate of substitution remains relatively constant across lineages.

The accuracy, or even the existence of the molecular clock has been an area of intense debate. Almost in parallel to the molecular clock hypothesis, Morris Goodman proposed the *generation-time effect hypothesis* (GOODMAN 1962; GOODMAN 1963). The generation-time effect hypothesis states that species with longer generation-time are expected to exhibit a slower rate of substitution compared to those with shorter generation time. This relationship between substitution rate and generation-time is thought to arise because of the idea that a major fraction of germline mutations originate from errors in DNA replication. Because species with longer generation-time undergo fewer numbers of replication in germ cells per unit time, fewer substitutions will accumulate. This hypothesis is strongly supported by molecular data. For example, analyses of non-coding (neutral) sequence data has revealed that the rate of substitution in old world monkeys is 30% faster than that of hominoids (YI *et al.* 2002). It has also been shown that the rate of substitution varies between closely related hominoids. The rate of substitution in the human lineage is slower than that of chimpanzee (ELANGO *et al.* 2006).

Intragenomic variation of single nucleotide substitution rate

In addition to the inter-specific variation in substitution rate discussed above, large variations of substitution rate exist within mammalian genomes. Interestingly, this variation occurs even at presumably neutral sites, suggesting that it is caused by mutation rate variation across the genome. Evidence for the existence of such variation was

initially reported using substitution rate at synonymous sites in protein coding regions (WOLFE *et al.* 1989). Several recent studies have confirmed this observation using non-coding sequence data from primates (ELLEGREN *et al.* 2003).

Variation of within-genome single nucleotide substitution rate has been studied at different scales (ELLEGREN *et al.* 2003). Mammalian substitution rate varies between-chromosomes, and within-chromosomes. Evidence for between-chromosome variation comes from several studies comparing substitution rate in primate autosomal non-coding regions across chromosomes [e.g., (CHEN *et al.* 2001)]. In addition, recent work has shown that the rate of substitution in the human Y chromosome is higher than that of other chromosomes (MAKOVA and LI 2002; MIKKELSEN T 2005). Assuming that most of the mutations occur because of errors during DNA replication, this observation has been attributed to the fact that 1) there are many more cell divisions in spermatogenesis than in oogenesis, and 2) Y chromosome is transmitted solely through males. There is a two-fold rate different between X and Y chromosomes, suggesting that the rate of mutation in males is ~5 times higher than that of females (MAKOVA and LI 2002). Such male mutation bias has also been observed in birds and fish (ELLEGREN and FRIDOLFSSON 1997; ELLEGREN and FRIDOLFSSON 2003).

A growing body of evidence now supports the between-chromosome variation of substitution rates. Substitutions in non-coding regions tend to cluster at a scale of 1-10 kb, which creates substantial variation of substitution rate within chromosomes (SILVA and KONDRASHOV 2002; SMITH *et al.* 2002). Interestingly, the rate of substitution in 5-kb blocks of non-coding DNA along the human lineage is strongly positively correlated to that that in the chimpanzee lineage (SMITH *et al.* 2002). This indicates that intra-

chromosomal variation is deterministic, repeatable, and may relate to intrinsic DNA sequence characteristics. Such a deterministic pattern has also been observed using human-mouse comparisons (WATERSTON *et al.* 2002).

At a finer scale, studies have shown that regional substitution rate variation within chromosome may be caused by hypermutability of certain nucleotides, which is greatly affected by the identity of neighboring nucleotides (sequence context). For example, it has been shown that transversion rates are biased towards G/C <-> A/T substitutions, suggesting that G/C nucleotides are more mutable (SMITH *et al.* 2002). Clustering analyses have revealed that the composition of neighbors (< 10 bps) plays an important role in mutagenesis (SILVA and KONDRASHOV 2002).

As explained above, one of the most important types of context dependent mutations is the DNA methylation dependent CpG mutation. CpG substitution rate has been shown to vary between regions of the genome. It has been shown that the rate of CpG substitution rate in 5' UTRs is lower compared to that of introns, exons, and intergenic regions (MAJEWSKI and OTT 2002). However, little work has been done to study the variation of CpG substitutions in neutral non-coding regions.

MOTIVATION

In spite of the importance of DNA methylation, and the prevalence of methylation dependent CpG substitutions in vertebrate genomes, its impact on genome evolution is under-explored. Several lines of thought indicate that DNA methylation must have affected both neutral and functional evolution of vertebrate genomes.

From the point of view of neutral evolution, first, CpG substitutions are expected to follow a more time-dependent molecular clock compared to substitutions at non-CpG, which are expected to follow a generation-time dependent molecular clock. This is because most mutations at non-CpG sites occur due to errors during DNA replication, whereas CpG mutations are not restricted to any particular stage of the cell cycle. Therefore, inter-specific variation of substitution rate must greatly depend on the relative contribution of CpG and non-CpG substitutions. Second, because deamination requires only a *local* strand separation, the rate of mutation at CpG sites must depend on the local G+C content. Moreover, the strength of this dependence must decrease with increasing distance from the CpG site. Given the prevalence of CpG substitutions in vertebrate genomes, this unique context dependence pattern must be a key factor affecting intra-genomic neutral substitution rate variation. However, none of the above hypotheses have been tested explicitly thus far.

From the functional evolutionary point of view, the presence of the LCG and the HCG classes of promoters of the human genome is a clear indication that DNA methylation plays a key role in gene regulation. However, the evolution of these classes of promoters is still unclear. Are the two classes of promoters a common feature of all globally methylated vertebrate genomes? Did they evolve as a consequence of the transition from mosaic to global DNA methylation pattern during early vertebrate evolution?

This dissertation focuses on the consequences of DNA methylation dependent single nucleotide mutations on vertebrate genome evolution. Following the ideas presented above, the topics investigated can be broadly grouped into two types.

1) Studies on neutral consequences

2) Studies on functional consequences

To study the neutral impact of DNA methylation (Chapter 2 and Chapter 3), we used non-coding sequence data from primate genomes. For investigating the functional impact of DNA methylation, we studied promoter regions from several distantly related vertebrate genomes (Chapter 4 and Chapter 5).

In Chapter 2, the inter-specific substitution rate variation between hominoids and old world monkeys is investigated separately for CpG sites and non-CpG sites. Results indicate that CpG sites follow a relatively time-dependent molecular clock, while non-CpG sites follow a generation-time dependent molecular clock. Thus, the genomic molecular clock as a whole is a mixture of both time-dependent and generation-time dependent molecular clocks. In other words primates exhibit a heterogeneous genomic molecular clock. A simple mathematical model was developed to elucidate the effect of CpG substitutions on inter-specific substitution rate variation, and the observations are discussed in the context of the controversy over the molecular clock hypothesis.

Chapter 3 investigates the intra-genomic variation of substitution rate at CpG dinucleotides in non-coding regions or primate genomes. The results of this study show that the rate of substitution at CpG dinucleotides is negatively correlated with local G+C content of the region. The strength of this correlation decays with increasing distance from the CpG site, and extends up to ~ 1500 – 2000 bps on each side of the site. This pattern is not observed in the case of mutations at GpC sites. We attribute this observation to the local DNA strand separation required for CpG mutations to occur. The

importance of such an observation is discussed in context of the molecular origins of mutations.

Chapter 4 investigates the functional impact of CpG substitutions in promoters of several distantly related vertebrate genomes (e.g., *Homo sapiens*, *Gallus gallus*, *Danio rerio*). Results indicate that vertebrate promoters fall into two structural classes – the hypermethylated low CpG content (LCG) and the hypomethylated high CpG content (HCG) promoters. This structural distinction is not observed in sea squirts, a close invertebrate outgroup. In all the vertebrate genomes analyzed LCG promoters are associated with tissue specific genes, whereas the HCG promoters are associated with broadly expressed genes. These observations suggest that the two classes of promoters evolved early in the vertebrate lineage, and the function of DNA methylation on gene regulation is conserved across diverse vertebrate taxa. A model for the evolution of the two structural classes of promoters is proposed and discussed.

In Chapter 5 promoter associated CpG islands, which are usually present in the HCG class of promoters, is analyzed in more detail with respect to the expression pattern of the associated genes. Results from this study show that the tissue-specificity of genes with promoters having a long CGI lies between those with short-CGI promoters (widely expressed) and those with no-CGI promoters (highly tissue-specific). These genes are highly conserved and involved in important biological processes like development, transcription regulation, and signaling. Long-CGI promoters also contain a larger number of RNA polymerase II binding sites, which are occupied in an intermediately tissue specific manner suggesting that long-CGIs are associated with genes that require more complex regulation strategies.

CHAPTER 2

HETEROGENEOUS GENOMIC MOLECULAR CLOCKS IN PRIMATES

ABSTRACT

Using data from primates, we show that molecular clocks in sites that have been part of a CpG dinucleotide in recent past (CpG sites) and non-CpG sites are of markedly different nature, reflecting differences in their molecular origins. Notably, single nucleotide substitutions at non-CpG sites show clear generation-time dependency, indicating that most of these substitutions occur by errors during DNA replication. On the other hand, substitutions at CpG sites occur relatively constantly over time, as expected from their primary origin due to methylation. Therefore, molecular clocks are heterogeneous even within a genome. Furthermore, we propose that varying frequencies of CpG dinucleotides in different genomic regions may have contributed significantly to conflicting earlier results on rate constancy of mammalian molecular clock. Our conclusion that different regions of genomes follow different molecular clocks should be considered when inferring divergence times using molecular data and in phylogenetic analysis.

INTRODUCTION

Organisms with longer generation-time tend to exhibit slower molecular clock than those with shorter generation-time, an effect known as "generation-time effect" (LAIRD *et al.* 1969; LI *et al.* 1996; YI *et al.* 2002). However, the extent (or even the

existence) of generation-time effect is of significant debate (KUMAR 2005; KUMAR and SUBRAMANIAN 2002). An opposing theory posits that molecular evolution occurs relatively constantly over time: in other words, molecular clocks are time dependent (EASTEAL and COLLET 1994; KUMAR 2005; KUMAR and SUBRAMANIAN 2002). Here we show that molecular evolution follows both generation-time–dependent and time-dependent molecular clocks, depending upon the molecular origins of the mutations considered.

A generation-time–dependent molecular clock implies that the majority of single nucleotide substitutions in germlines arise from errors during DNA replication (GOODMAN 1962; GOODMAN 1963). However, some mutations may occur independently from DNA replication. This is especially pertinent for transitions at CpG dinucleotides (henceforth, CpG substitutions). CpG substitutions are the most frequent single nucleotide substitutions in vertebrate genomes, accounting for more than a quarter of all substitutions between the genomes of human and chimpanzee (MIKKELSEN T 2005; NACHMAN and CROWELL 2000). Naturally, they play critical roles in several key genetic mechanisms and disease (JONES and LAIRD 1999; MEUNIER *et al.* 2005; ROBERTSON and WOLFFE 2000).

CpG dinucleotides are hypermutable because the cytosines in CpG dinucleotides are targets of DNA methylation in vertebrate genomes (BIRD 1980). Methylated cytosine rapidly mutates to thymine via spontaneous deamination, causing a C to T (G to A in the complementary strand) transition (BIRD 1980; DUNCAN and MILLER 1980). While DNA replication occurs in a specialized stage of the cell cycle, methylation is not confined to replicating DNA: germline cells are methylated early in their development and stay

methylated until global de-methylation occurs after fertilization (BIRD 2002; LI 2002). Therefore, methylation-origin mutations will accumulate in a rate proportional to the total amount of time germ cells are methylated between generations. In other words, the molecular clock at CpG dinucleotides should be relatively constant over time.

Indeed, statistical inferences using approximately 2 Mbps of sequence data have suggested that CpG substitutions follow relatively constant molecular clock in mammals (HWANG and GREEN 2004). In addition, a recent analysis of male mutation bias in humans and chimpanzees have shown that CpG dinucleotides exhibit much lower male mutation bias than other sites (TAYLOR *et al.* 2006). Since male-mutation bias is caused by the more frequent DNA replications in male germlines compared to female germlines (LI *et al.* 2002), the finding that there is lower male mutation bias in CpG dinucleotides is consistent with the idea that CpG substitutions follow a relatively time-dependent molecular clock.

In this paper, we sought to directly compare genomic molecular clocks of CpG dinucleotides and other sites. To achieve this goal, we focused on catarrhines, specifically two hominoid species (human and chimpanzee) and two Old World monkeys (rhesus macaque and baboon). These four species are chosen because they satisfy two criteria. First, because these species are closely related, we can identify sites that have been part of a CpG dinucleotide in recent past (CpG sites) and other sites with high confidence (MEUNIER and DURET 2004). Second, hominoids and Old World monkeys have markedly differently generation times. According to (GAGE 1998), average generation times in Old World monkeys is 11.4 years, while in chimpanzees and humans, they are 22 and 28 years, respectively. As a consequence of the difference in generation times, evolutionary

rates of replication-dependent substitutions are slower in hominoids than in Old World monkeys (GOODMAN 1962; LI *et al.* 1996; YI *et al.* 2002).

Utilizing genomic data from these species, we demonstrate that indeed CpG substitutions exhibit a relatively time-dependent molecular clock, in contrast to generation-time–dependent genomic molecular clock. Furthermore, we propose that heterogeneous molecular clocks among different genomic regions may have contributed to conflicting earlier results on the degree of generation time effect in mammals.

RESULTS AND DISCUSSION

Slower Molecular Evolution of Hominoid Genomes than Old World Monkey Genomes

We first re-evaluated the difference in evolutionary rates between hominoids and Old World monkeys. We analyzed approximately 28 Mbps of genomic sequence alignments to compare rates in human (a hominoid) and baboon (an Old World monkey) using a relative rate test (WU and LI 1985; YI *et al.* 2002). Sequence data from marmoset (a New World monkey) was used as an outgroup. We found that rates in humans are on average 28.4% slower than those in baboons in introns and intergenic regions (Table 2.1, $p < 0.001$), confirming earlier results (LI *et al.* 1996; STEIPER *et al.* 2004; YI *et al.* 2002). Because data used in this analysis account for approximately 1% of the human genome and from several different chromosomes, we can conclude that the canonical genomic molecular clocks in primates exhibit significant generation-time effect.

Table 2.1. Hominoid-Rate Slowdown Tested Using Genomic Sequence Data from Human, Baboon, and a Marmoset

	Ratio of Old World Monkey Branch to Hominoid		
	All Sites	CpG Sites	Non-CpG Sites
Transitions	1.27*	1.05^{NS}	1.33*
Transversions	1.31*	0.99^{NS}	1.29*
All substitutions	1.28*	1.03^{NS}	1.32*

* $p < 0.001$ by relative-rate test.
^{NS} Not Significant

We also constructed a five-species phylogeny of human, chimpanzee, baboon, macaque, and marmoset using data for 1.9 Mbps of sequences orthologous to the human chromosome 7 (hg17.chr7: 115404472-117281897 ;ENCODE region ENm001). High-quality sequence data are available for all five species analyzed in this study (accession numbers NT_086357.2, NT_165329.1, NT_086378.3 and NT_165339.1, NT_086504.2 for human, chimpanzee, baboon, macaque, and marmoset, respectively). Figure 2.1 shows a Neighbor-Joining tree (SAITOU and NEI 1987) of the five species. Focusing on the ancestral hominoid and ancestral Old World monkey branches, the ratio of the number of substitutions in the Old World monkey branch to the hominoid branch is approximately 1.36, similar to the values estimated from the comparison between the human and baboon genomes. These results confirm the "hominoid rate slowdown" theory proposed more than 40 years ago (GOODMAN 1962).

Our next goal was to compare the molecular clocks at CpG and non-CpG sites separately. However, because of the difficulty in correcting for multiple hits, we cannot

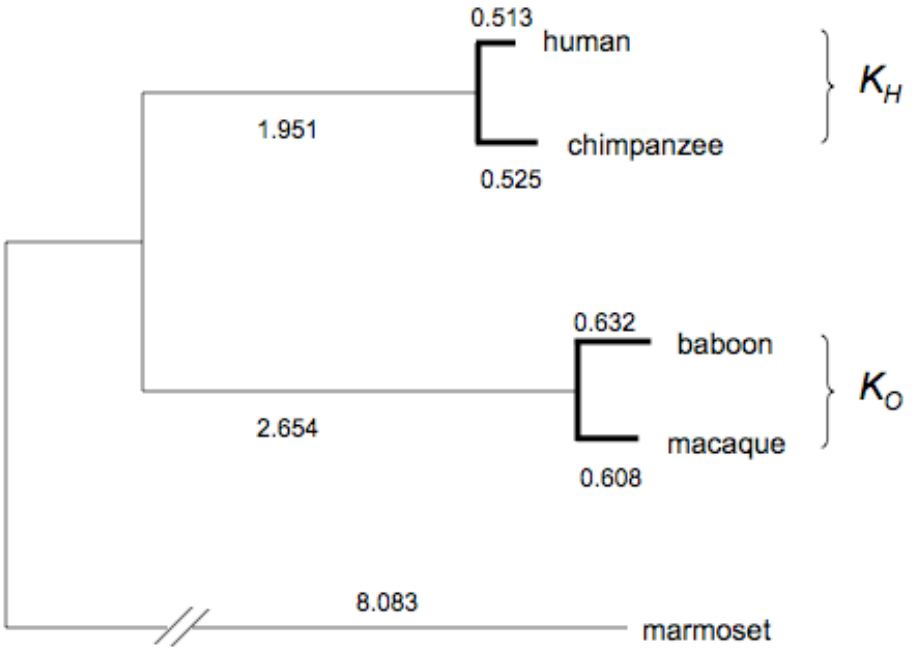


Figure 2.1. A Neighbor-Joining Tree of Five Primate Species, Generated Using High-Quality Data from the Encode Region ENm001.

The numbers of substitutions per 100 sites in each branch, using the 2-parameter correction. are shown.

easily analyze substitutions at CpG sites in this phylogenetic setting. Therefore, we proceeded to use data only in catarrhines, where we can accurately infer rates in CpG and non-CpG sites (MEUNIER and DURET 2004; MEUNIER *et al.* 2005; TAYLOR *et al.* 2006).

Different Molecular Clocks of CpG Sites and Non-CpG Sites

We constructed four-species alignments of two hominoids (human and chimpanzee) and two Old World monkeys (rhesus macaque and baboon) (Figure 2.2). These species pairs provide a unique opportunity to study time-dependent and generation-time–dependent clocks. Critical to our work, the divergence time between the hominoid pair is similar to that of the Old World monkey pair (BRUNET *et al.* 2002; STEIPER *et al.* 2004). The split between human and chimpanzee is estimated to be 6 to 8 million years ago (Mya), based upon fossil records. In particular, the earliest fossil hominin, *Sahelanthropus tchadensis*, has been dated to late Miocene, at least 7 Mya (BRUNET *et al.* 2005; BRUNET *et al.* 2002). The split between rhesus macaque and baboon is calibrated by using an estimate for the split between macaques and papionins. The earliest fossil evidence of papionins is dated to be 6 to 8 Mya (BRUNET *et al.* 2002; STEIPER *et al.* 2004). Therefore, divergence times of the two species within each pair are similar. In other words, $T_O/T_H \approx 1$ (Figure 2.2). In contrast to this similarity of within-pair divergence times, evolutionary rates are known to differ between these two groups: as explained in the introduction and demonstrated above, genomic evolutionary rates in hominoids are slower than rates in Old World monkeys.

We have two contrasting predictions for a time-dependent versus a generation-time–dependent molecular clock. For replication origin (hence generation-time dependent) mutations, the pairwise sequence divergence in the Old World monkey pair

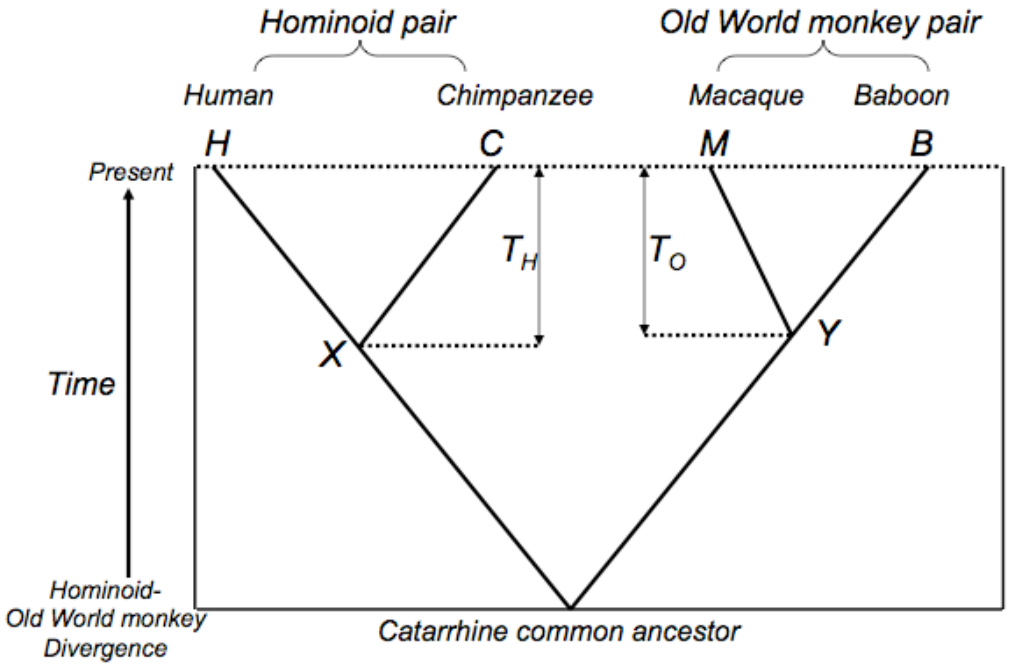


Figure 2.2. Phylogeny of the Four Taxa Analyzed in This Study.

T_O denotes the time since the split between the two Old World monkey species, and T_H denotes the time since the split between the two hominoids. Fossil records suggest that T_O and T_H are very close to each other. X and Y denote the common ancestors of human-chimpanzee and of macaque-baboon, respectively. The genetic divergence between the two hominoid species (K_H) is the sum of K_{HX} and K_{HC}. Likewise, K_O is the sum of K_{MY} and K_{BY}.

($K_O = K_{MY} + K_{BY}$ in Figure 2.2) should be greater than the pairwise sequence divergence in the hominoid pair ($K_H = K_{HX} + K_{CX}$ in Figure 2.2). On the other hand, a time-dependent molecular clock predicts that K_O is similar to K_H.

We examined the molecular clocks in CpG and non-CpG sites separately (see Materials and Methods). To directly compare mutations caused by deamination of methylated cytosines to other transitions occurring during replication, we first analyzed only C-to-T (and G-to-A) transitions. A distinctive pattern emerged: K_O/K_H is 1.03 in CpG sites (95% confidence interval [CI], 0.92 to 1.15), while it is 1.31 in non-CpG sites (95% CI, 1.25 to 1.37). These two types of sites clearly harbor different molecular clocks. Similar trends were discovered when introns and intergenic regions are considered separately, or when repetitive and nonrepetitive sequences are compared separately (Figure 2.3).

We then considered all single nucleotide substitutions that occurred in CpG and non-CpG sites and found the same pattern. The ratio K_O/K_H in non-CpG sites is 1.18 (95% CI, 1.15 to 1.22). In comparison, in CpG sites, K_O/K_H is 1.00 (95% CI, 0.89 to 1.11). Again, the results are similar when introns and intergenic regions are considered separately, or when repetitive and nonrepetitive sequences are compared separately.

Because human-chimpanzee (hominoid pair) and rhesus macaque-baboon (Old World monkey pair) are extremely closely related, estimates of pairwise sequence divergence are affected by common ancestral polymorphism (EBERSBERGER *et al.* 2002; ELANGO *et al.* 2006; MAKOVA and LI 2002). The common ancestor of the human and chimpanzee is thought to have much larger effective population size than the current human population (CHEN and LI 2001; WALL 2003). Rhesus macaque and baboon also

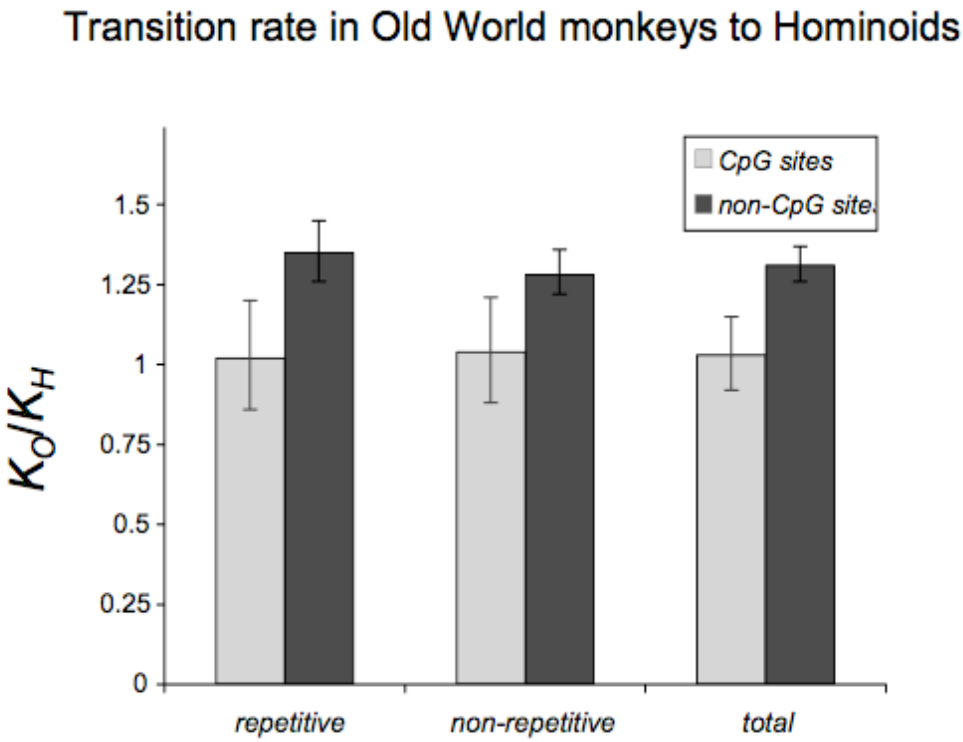


Figure 2.3. Contrasting Molecular Clocks of Transitions at CpG Sites versus Those at Non-CpG Sites

The Y-axis shows the rate difference in the baboon-macaque pair to that in the human-chimpanzee pair. The Old World monkey pair has accumulated significantly more transitions in non-CpG sites, as expected by the generation time effect. In contrast, transitions at CpG sites, which are primarily of methylation-origin, show no difference between the two pairs. Data are shown for all sites, repetitive sites (as identified from the RepeatMasker program), and non-repetitive sites (after removing repetitive sites). Confidence intervals are generated by bootstrapping 10,000 times.

harbor comparable levels of genetic diversity to hominoids. For example, (WALL *et al.* 2003) estimated a nucleotide diversity of 0.13% in a noncoding region of rhesus macaques.

Such substantial ancestral polymorphism will effectively reduce the observed rate difference between hominoid and Old World monkey pair: the observed pairwise divergence between rhesus macaque and baboon (K_O) is the sum of ancestral diversity (π_Y, see Figure 2.2) and the fixed difference between rhesus macaque and baboon (denoted as P_O). Likewise, the pairwise divergence between human and chimpanzee, $K_H = \pi_X + P_H$. We are interested in the ratio P_O/P_H while we only have access to K_O/K_H. When comparing distantly related species, the level of ancestral diversity is negligible relative to the fixed difference. However, between closely related species such as human-chimpanzee and macaque-baboon, ancestral diversity is substantial compared to the fixed difference. For example, π_X can be as much as $\frac{1}{2}P_H$ (CHEN and LI 2001). Therefore, K_O/K_H will underestimate P_O/P_H.

To address this concern, we used the estimates obtained for CpG and non-CpG sites in hominoids (TAYLOR *et al.* 2006) to correct for the effect of ancestral polymorphism. After such corrections, K_O/K_H for non-CpG sites is 1.18 to 1.26 (Table 2.2). In contrast, in CpG sites, K_O/K_H is close to 1.00 even after correcting for the effect of ancestral polymorphism using estimates for CpG sites (Table 2.2). However, these values should be taken with caution, given the uncertainties associated with ancestral diversity as well as with divergence time estimated from fossil records.

Table 2.2. The Ratio of the Pairwise Divergence between Macaque and Baboon *(K_O)* to the Pairwise Divergence between Human and Chimpanzee *(K_H)*, Using All Substitutions. We used the estimates nucleotide diversity (π) for CpG and non-CpG sites separately obtained from the human genome to correct for the effect of ancestral polymorphism, for several different scenarios. Each row represents corrections using different level of ancestral polymorphism. Confidence intervals are obtained by bootstrapping 10,000 times.

	K_O/K_H (95% CI)	
Levels of ancestral polymorphism	CpG Sites	Non-CpG Sites
No correction	0.998 (0.892-1.119)	1.178 (1.147-1.210)
π	0.998 (0.879-1.135)	1.192 (1.159-1.227)
2π	0.998 (0.864-1.155)	1.210 (1.173-1.248)
4π	0.997 (0.813-1.216)	1.255 (1.220-1.303)

For completeness, we also analyzed the rate difference for CpG and non-CpG sites using the above three-species alignment (human, baboon, and marmoset). Even though this comparison is less reliable due to the difficulty in correcting for multiple hits (see above), we obtained similar results. We observe that the non-CpG sites (the majority of sites) show substantial rate difference between the human and the baboon genomes. In contrast, CpG sites show little difference in evolutionary rates between hominoid and Old World monkeys.

In summary, CpG and non-CpG sites show statistically different molecular clocks in various phylogenetic comparisons, indicating that the difference in two types of molecular clocks is a salient picture of molecular evolution in primate genomes.

Factors that May Affect K_O/K_H for CpG and Non-CpG Sites

Here we review some of the potential factors that can affect our conclusions. An important assumption in our work is that the divergence time between the hominoid pair is similar to that of the Old World monkey pair. This was mainly based upon fossil records (BRUNET *et al.* 2002; STEIPER *et al.* 2004). However, because fossil records are inherently associated with large variance in dates, let us consider the inference from molecular data.

If we measure the divergence between the Old World monkey pair to that between the hominoid pair in the five species phylogeny shown in Figure 2.1 (equivalent to K_O/K_H in Figure 2.2), it is 1.2. This is different from the ratio obtained from the comparison of the ancestral Old World monkey branch to the hominoid branch, which

was 1.36. The discrepancy between these two estimates can be explained by at least two mechanisms, which are not mutually exclusive of each other.

First, as mentioned earlier, estimating evolutionary rates between closely related species, such as human-chimpanzee and macaque-baboon, is significantly affected by ancestral polymorphism (EBERSBERGER *et al.* 2002; ELANGO *et al.* 2006; MAKOVA and LI 2002). If we use estimates of the ancestral polymorphism in hominoids (CHEN and LI 2001; WALL 2003) to correct for the effect of ancestral polymorphism, the ratio of K_O/K_H increases, close to the value estimated from the ancestral branch. For example, if we assume that the average nucleotide diversities of the ancestral Old World monkey and hominoid populations were 0.4%, the corrected ratio of K_O/K_H increases to 1.32.

The second possibility is that the actual time in the Old World monkey pair (T_O) is slightly shorter than the time in the hominoid pair (T_H). Because fossil records provide only the "minimum" divergence time between lineages, the actual divergence time can differ significantly, and the divergence of human and chimpanzee may have occurred before the divergence of macaque and baboon. Therefore, K_O/K_H will underestimate the true rate difference. According to this possibility, the CpG clock in our data also underestimates the actual rate difference, indicating that some fraction of CpG substitutions follows a generation-time–dependent molecular clock. We believe that this scenario at least partially explains the observed discrepancy, because some substitutions at CpG sites occur during replication. This interpretation is also in accord with the weak but still significant male mutation bias in hominoids (TAYLOR *et al.* 2006).

Our study uncovered significant heterogeneity in the degree of generation time effect among different types of single nucleotide substitutions. In particular, when

substitutions are divided into transitions and transversions, the latter exhibited less generation-time effect than transitions. In fact, in CpG sites, there were more transversions in the human-chimpanzee pair than in the baboon-macaque pair (58 versus 39). However, the numbers are rather small (since most substitutions at CpG sites are transitions due to methylation) so it is not clear whether this reflects a true underlying pattern. In non-CpG sites, the ratio K_O/K_H estimated from transitions was 1.31 while the ratio from transversions was 1.14 (the overall ratio was 1.18). Whether this discrepancy reflects differences in molecular mechanisms between transitions and transversions is an interesting question and should be pursued further.

Effect of CpG Dinucleotides on Hominoid Rate Slowdown and Mammalian Molecular Clock

Our findings shed important light on the controversy over mammalian molecular clock. Generation-time effect was clearly demonstrated when closely related species were compared or when noncoding sequences were used (HWANG and GREEN 2004; STEIPER *et al.* 2004). However, among relatively distant mammalian species, weak generation-time effect was observed (KUMAR and SUBRAMANIAN 2002; WU and LI 1985). Note that due to sequence availability and alignability, synonymous sites were often used when comparing distantly related species.

We propose that varying proportions of CpG dinucleotides in different data sources can contribute to conflicting conclusions on the nature of genomic molecular clocks. Three observations led to this hypothesis. First, CpG molecular clock runs much faster than clocks at other sites, at least in primates. Assuming that human and chimpanzee diverged 7 Mya (BRUNET *et al.* 2002), we estimate that CpG sites and non-

CpG sites undergo single nucleotide substitutions at a rate of 1.03×10^{-8} per site per year and 0.68×10^{-9} per site per year, respectively, from our data. Second, molecular clocks at CpG sites are relatively constant over time. Third, the proportion of CpG dinucleotides is heterogeneous among different genomic regions (SUBRAMANIAN and KUMAR 2003). In particular, 4-fold degenerate sites are enriched with CpG sites, over 10% (SUBRAMANIAN and KUMAR 2003), while noncoding regions have less than 3% CpG dinucleotides (SUBRAMANIAN and KUMAR 2003; TAYLOR *et al.* 2006). Hence, molecular clocks in regions with relatively abundant CpG sites (such as 4-fold degenerate sites) may be dominated by the rapid and time-dependent CpG clock, while regions relatively devoid of CpG sites (such as noncoding regions) follow generation-time–dependent molecular clock.

To investigate this prediction, we compared results from different studies in Table 2.3, focusing on two comparisons: between hominoids and Old World monkeys (hominoid rate slowdown), and between primates and rodents. Note that earlier studies on molecular clock did not consider CpG content as a determinant of molecular clock. Therefore, they did not investigate the effect of CpG content on molecular clock. Because some studies used noncoding regions while others used 4-fourfold degenerate sites, different studies analyzed different data in relation to CpG content (Table 2.3). We did not include the results from (KUMAR and SUBRAMANIAN 2002) in this table, because they removed a substantial amount of data that did not pass the "homogeneity test," and the relationship between this test and CpG dinucleotide content is not clear. For example, they discarded 46% of the data in their human-mouse comparison (KUMAR and SUBRAMANIAN 2002).

We can now compare how the data in Table 2.3 fit our hypothesis. First, when we compare results from all sites, the rate difference between lineages is greater in noncoding regions than in 4-fold degenerate sites. Moreover, in noncoding regions, the rate difference for CpG sites is lower than for all sites or non-CpG sites. Similarly, in 4-fold degenerate sites, the rate difference in non-CpG sites is higher than in all sites. These trends support our hypothesis.

Since we have reasonable estimates of CpG and non-CpG rates in primates (see above), we can investigate how well our hypothesis fits the data in detail. The number of substitutions in hominoids since the split from Old World monkeys can be approximated as $\left(pk_{CpG} + (1-p)k_{non-CpG}\right)T$, where p is the proportion of CpG sites, k_{CpG} and $k_{non\text{-}CpG}$ represent substitution rates per site per year in CpG sites and non-CpG sites, respectively, and T is the time since the split. The observed ratio of Old World monkey branch to hominoid branch can then be expressed as $\frac{pk_{CpG} + r(1-p)k_{non-CpG}}{pk_{CpG} + (1-p)k_{non-CpG}}$, where r represents the ratio of the branch lengths determined by the generation-time dependent molecular clock. Figure 2.4 shows this ratio as a function of p, using the rates inferred from our data. In case when $r = 1.4$, the observed ratios from regions with 12% and 2.5% CpG dinucleotides (analogous to 4-fold degenerate sites and intergenic regions) are 1.12 and 1.29, respectively.

We compared these theoretical expectations to observed values by analyzing rates between hominoids and Old World monkeys in 4-fold degenerate sites, from 41 autosomal genes (Table S2). The proportion of 4-fold degenerate sites that belong to CpG dinucleotides in any of the three species compared in this dataset is 11.0%. This is likely

Table 2.3. Rate Differences between Lineages from Various Data Sources.
Noncoding regions usually have low CpG content (typically less than 3%, see (SUBRAMANIAN and KUMAR 2003) for example and similar proportions were found in our data), while fourfold degenerate sites are enriched with CpG sites (more than 10%, and similar proportions were found in our data). Therefore, molecular clock in fourfold degenerate sites may appear more time-dependent than that in noncoding regions. According to this prediction, the rate difference is greater in noncoding regions than in fourfold degenerate sites. CpG sites in noncoding regions show lower rate difference than all sites or non-CpG sites. Similarly, in fourfold degenerate sites, the rate difference increases when only non-CpG sites are used. We also performed additional analyses using fourfold degenerate sites from mammals and report the results.

		CpG Sites	All Sites	Non-CpG Sites
Ratio of Old World Monkey branch to hominoid branch	Noncoding regions	1.01*	1.25*	
			1.33	
				1.45
		1.03	1.28	1.31
	4-fold degenerate sites		1.09	1.27
Ratio of rodent branch to primate branch	Noncoding regions	1.68*	2.87*	
	4-fold degenerate sites		1.81	
			2.57†	2.87†

***** In order to calculate rate difference for data from (HWANG and GREEN 2004), human and chimpanzee branch lengths were averaged to estimate hominoid branch length, whereas baboon and macaque branch lengths were averaged to estimate Old World monkey branch length. For the primate-rodent comparison, rat and mouse branch lengths were averaged to estimate rodent branch length. Data for all sites came from the phylogenetic tree in Supporting Figure 11 in (HWANG and GREEN 2004). Data for CpG sites came from the phylogenetic tree in Supporting Figure 22 (HWANG and GREEN 2004), which describes NCG→T mutations (i.e., CpG → TpG mutations).

† In this comparison, because of the long divergence time, our definition of non-CpG sites may not be effective in removing all sites that have been a part of CpG dinucleotides. Despite such limitations, we observe that the rate difference increases when we use only non-CpG sites. CpG sites cannot be accurately identified in this comparison due to the long divergence time.

an underestimate of the true proportions of sites that have been part of a CpG dinucleotide, since the divergence time between the three species is rather long. The ratio of the Old World monkey branch to the hominoid branch was 1.09 when all sites were used (Table 2.3). When we removed CpG-prone sites (sites preceded by C or followed by G, as used in (KEIGHTLEY *et al.* 2005; MEUNIER and DURET 2004; MEUNIER *et al.* 2005) from the 4-fold degenerate sites, the aforementioned ratio was increased to 1.27 (Table 2.3). Recall, when only noncoding sites were used, this ratio was 1.28 (Table 2.1), which increased to 1.31 when we removed CpG sites. The proportion of sites that belong to CpG dinucleotides in noncoding sites in our data is 2.5%. Therefore, these values are in excellent accord with the above-mentioned model.

It should be noted, however, that the above model ignores other factors that affect regional mutation rate variation, such as GC content and recombination (HELLMANN *et al.* 2003; YI *et al.* 2002). Also, as discussed above, different mutations (such as transitions and transversions) may have different substitution rates between lineages. Hence, partitioning rates into only two categories is likely to be a simplification. Furthermore, identifying sites that have been part of a CpG dinucleotide in the past is a challenging problem (SIEPEL and HAUSSLER 2004). Lineage-specific rates are also affected by ancestral generation times and effective population sizes. Further studies are necessary to determine the roles of generation-time–dependent and time-dependent molecular clocks on genome evolution.

Nevertheless, it is clear that the heterogeneity of molecular clocks due to different mutational origins can significantly alter rate differences between taxa. This effect should

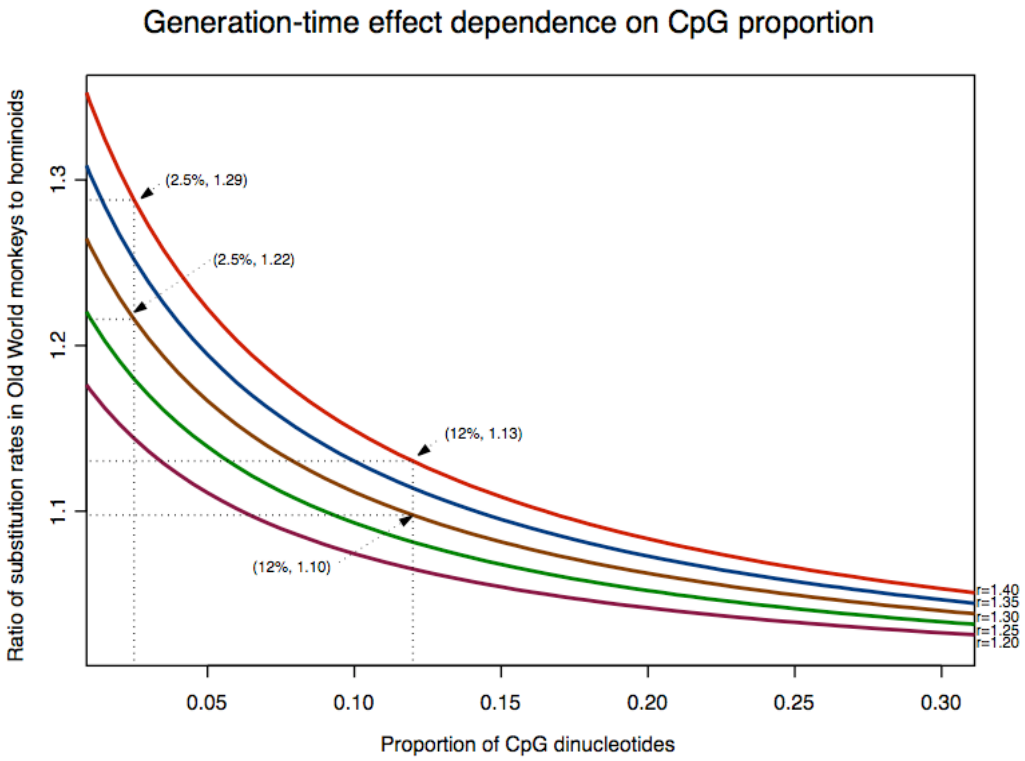


Figure 2.4. The Proportion of CpG Sites in Data Affects the Degree of Hominoid Rate Slowdown

We considered a simple model in which all sites can be classified into either CpG sites or non-CpG sites, and estimated evolutionary rates in hominoids from the human-chimpanzee comparison. The X-axis is the proportion of CpG sites in the data. The Y-axis is the observed degree of hominoid rate slowdown, shown as the ratio of the substitution rate in Old World monkeys to the rate in hominoids, given the 'true' ratio (determined by the generation-time effect), depicted as *r*. While regions relatively devoid of CpG sites will reflect the true generation-time effect, the observed ratio approaches 1 as the data include more CpG sites (i.e., the substitution rate in hominoids and Old World monkeys will be similar). Data points for when data consists of 2.5% and 12% CpG sites for $r = 1.3$ and 1.4 are shown for convenience.

be taken into account when molecular clocks are used to infer divergence times and to reconstruct phylogenetic history.

MATERIALS AND METHODS

Noncoding data mining and assembly

Because accurate identification of CpG sites is critical in our analyses, we used two precautions. First, we analyzed sequences between closely related primates only. Earlier studies have shown that within catarrhines (hominoids and Old World monkeys), we can directly derive rates of CpG substitutions using comparative methods. Specifically, we can confidently determine "CpG sites" (sites for which the ancestral state was part of a CpG) and extract rates of CpG substitutions using parsimony (MEUNIER and DURET 2004; MEUNIER *et al.* 2005; TAYLOR *et al.* 2006). Moreover, we can also identify sites that have not been a part of CpG dinucleotides (non-CpG sites), to be used as a control for replication-origin substitutions (MEUNIER and DURET 2004; MEUNIER *et al.* 2005; TAYLOR *et al.* 2006). Second, we only used high-quality sequence data, because data obtained from whole genome assemblies include errors in sequencing and assembly that can cause erroneous conclusions regarding rate difference between lineages (ELANGO *et al.* 2006).

For the human-baboon-marmoset dataset, we obtained approximately 28 Mbps of high-quality data (BAC-based) from the ENCODE project (ENCODE-CONSORTIUM 2004).

For the human-chimpanzee-baboon-macaque (HCBM) dataset, we mined high-quality BAC-based sequences from GenBank. The HCBM dataset consists of BAC-based

sequence data orthologous to human chromosome 7 (hg17.chr7:114505472-117281897; Encode region ENm001). This is obtained by aligning NT_086357.2 (LANDER *et al.* 2001), NT_165329.1 (chimpanzee), NT_086378.3 (baboon), and NT_165339.1 (macaque) sequences.

We assembled additional orthologous alignments among the four species using the following procedure. First, we searched the GenBank database for sequences from baboon (*Papio anubis* or *Papio hamadryas*), macaque (*Macaca mulatta*) and chimpanzee (*Pan troglodytes*) BAC clones. We obtained sequence data for 377, 276, and 1641 BACs from baboon, macaque, and chimpanzee, respectively. Next, we identified orthologous BAC clones among these species, using BLAST (ALTSCHUL *et al.* 1990) and other methods as in (KENT *et al.* 2003). We found 25 baboon BAC clones that had both macaque and chimpanzee orthologs. We then localized orthologous human region for each of these 35 orthologous clones using BLAT (KENT 2002). We reconfirmed the orthology between baboon, chimpanzee and macaque BAC clones by ensuring that the regions where these BAC clones independently map to the human genome overlap with each other. We then removed the BAC clones overlapping with ENm001. Finally, we removed sequences from sex chromosomes. As a result, we obtained 16 genomic regions, shown in Table S1.

Analysis of fourfold degenerate sites

Primate Comparison: All sequence data for the primate fourfold degenerate site comparisons was downloaded from GenBank (BENSON *et al.* 2006). Accession numbers for all genes used in primate comparison is available in Table S2. A portion of the homologous genes in this dataset was also identified via the HOVERGEN database

(DURET *et al.* 1994). Sequence data for the human-mouse-dog comparison was downloaded from the Ensembl database (BIRNEY *et al.* 2006). Any genes that underwent recent gene duplications or did not meet the stringent minimum length of 445 nucleotides were removed from the dataset. Sequences were aligned using CLUSTALW via a BioPerl package (STAJICH *et al.* 2002). After alignment of homologous genes, any genes containing lineages with a negative K_4 value were removed from the dataset.

Primate-Rodent Comparison: Known genes from human, mouse and dog were downloaded from Ensembl (BIRNEY *et al.* 2006). To find orthologous sequences, we used the OrthoMCL algorithm (LI *et al.* 2003), which uses an all-to-all BLASTP results to generate a graph of orthologs and paralogs. We used default parameters except for E-value < 10^{-10} to ensure orthology. As a result, we constructed 3,494 orthologous gene trios among the three species. Next steps were performed as described in the primate comparison described above.

Sequence curation, data annotation, and statistical analyses

CpG Islands: CpG islands were identified using the algorithm by Takai and Jones (TAKAI and JONES 2002) with the following conditions: GC content > 55%, Observed/Expected CpG contents > 0.65, length = 200. Since the majority of CpG islands are hypomethylated and do not reflect substitutions of methylation-origin, we removed them from further analysis.

Sequence Annotation: Repetitive elements were annotated using the RepeatMasker program. Noncoding regions are identified as in (ELANGO *et al.* 2006).

Relative Rate Test: The two-parameter model (KIMURA 1980) was used to correct for multiple hits. We used a relative rate test (WU and LI 1985; YI *et al.* 2002) to test for rate difference between hominoid and Old World monkeys using New World monkey species as outgroup (Table S2). To compare rate difference between human and mouse, we used dog as an outgroup.

Classification and Rate Estimation of CpG sites and non-CpG sites: We used the method in (MEUNIER *et al.* 2005) to identify CpG and non-CpG sites. Specifically, CpG sites are defined as the middle base of the following patterns: XNG/XCG/XCG/XCG, with X denoting any nucleotide except C to avoid overlapping CpGs. N can occur in any of the four sequences. Sites fitting the complementary pattern (CGY/CGY/CGY/CNY, Y not G) are also considered as CpG sites. As a control, sites expected to have never been part of a CpG dinucleotides since the last common ancestor of the four species ('non-CpG sites') are defined as sites not preceded by C nor followed by G (MEUNIER and DURET 2004; MEUNIER *et al.* 2005). Sites that do not satisfy either classification are defined as 'ambiguous sites' and excluded from the analysis. A simulation study has shown that this classifying scheme can accurately identify CpG sites and non-CpG sites in catarrhines (MEUNIER and DURET 2004). Substitutions are then inferred using unweighted parsimony using only such sites. Confidence intervals for estimated rates are derived from bootstrapping 10,000 times.

CHAPTER 3

MUTATIONS OF DIFFERENT MOLECULAR ORIGINS EXHIBIT CONTRASTING PATTERNS OF REGIONAL SUBSTITUTION RATE VARIATION

ABSTRACT

Transitions at CpG dinucleotides, referred to as 'CpG substitutions', are a major mutational input into vertebrate genome evolution and a leading cause of human genetic disease. The prevalence of CpG substitutions is due to its mutational origin, which is dependent on the DNA methylation process. In comparison, other single nucleotide substitutions (for example those occurring at GpC dinucleotides) are believed to arise mainly from errors during DNA replication.

Despite the prevalence and importance of CpG substitution, relatively little is known on the patterns and causes of CpG substitution rate variation. Here we analyzed high quality BAC-based data from human, chimpanzee, and baboon to investigate regional variation of CpG substitution rates.

We show that CpG substitutions exhibit substantial regional variation within primate genomes, and that their pattern is consistent with differences in methylation level and susceptibility to subsequent deamination. In particular, we propose a novel hypothesis, referred to as the 'distance-decaying' hypothesis, positing that due to the molecular mechanism of a CpG substitution, their rates are correlated with the stability of

double stranded DNA surrounding each CpG dinucleotide, and the effect of local DNA stability may decrease with distance from the CpG dinucleotide.

Consistent with our 'distance decaying' hypothesis, rates of CpG substitution are strongly (negatively) correlated with *local* G+C content, where the local influence extends to 1500-2000 bps. Moreover, the influence of G+C content decays as the distance from the target CpG site increases. GpC sites, in contrast, do not exhibit such 'distance-decaying' relationship. Our results highlight an example of the distinctive properties exhibited by methylation-dependent substitutions versus substitutions mostly arising from errors during DNA replication.

INTRODUCTION

Elucidating the causes of regional variation of substitution rates is a prominent topic in molecular evolutionary studies (CASANE *et al.* 1997; GAFFNEY and KEIGHTLEY 2005; HELLMANN *et al.* 2005; HWANG and GREEN 2004; MIKKELSEN T 2005; WOLFE *et al.* 1989). While most studies consider all types of point mutations together, different types of mutations arise from distinctive molecular processes. In this paper, we contrasted regional variations of two types of mutations that are prevalent in mammalian genomes according to their molecular origins.

The first type is transitions at CpG dinucleotides, which is caused primarily by methylation of cytosine (BIRD 1980; SVED and BIRD 1990). Methylation followed by deamination causes a C to T transition (or G to A transition in the complementary strand) at a CpG dinucleotide. We refer to C to T or G to A substitutions at a CpG site as 'CpG substitutions'. These are the most frequent point mutation in the human genome (ARNDT

et al. 2003; MIKKELSEN T 2005), and often the basis of human genetic disease (KESHET *et al.* 2006; LI *et al.* 2002). In contrast, most of other point mutations are believed to occur from errors in DNA replication. Because of the differences in mutational mechanisms, these two types of mutations may behave differently in their relationship to regional sequence context.

In particular, efficiency of the deamination step may affect regional variation of CpG substitutions (FRYXELL and ZUCKERKANDL 2000). A prerequisite for the deamination process is the insertion of a water molecule between the DNA strands, via temporary '*melting*' or strand separation, which requires thermodynamic energy (FREDERICO *et al.* 1990; FREDERICO *et al.* 1993; FRYXELL and ZUCKERKANDL 2000). Bonds between G and C nucleotides require more thermodynamic energy to break, compared with A and T nucleotides. Therefore, the substitution rate at CpG dinucleotides may be negatively correlated with GC content (defined as the percentage of G and C nucleotides). Recently, several studies have shown that this prediction is true in mammalian genomes (ARNDT *et al.* 2005; FRYXELL and MOON 2005).

However, several other types of single nucleotide substitution also exhibit negative correlation with the G+C content (ARNDT *et al.* 2005). Therefore, the observed negative relationship may be caused by a more general molecular mechanism, such as biased gene conversion. Furthermore, it is unlikely that the G+C content of the whole sequence segment affects the probability of a CpG site to mutate. Rather, the effect of G+C content on CpG substitution may be confined to sites that are nearby the CpG site. In other words, the effect of sequence context on a CpG substitution is likely to be *local*.

Based upon this logic, we predict that G+C content has a strong *distance decaying* influence on the CpG substitution rate, because only *local DNA melting* is required for deamination to occur. In particular, the G+C content near the target CpG site will have a strong effect on the rate of CpG substitution. This effect will decay as the distance to the target CpG site increases. We present evidence that is consistent with our *distance decaying local GC influence hypothesis*.

In addition to our main results on the distance decaying relationship, we discuss other causative factors of regional heterogeneity of CpG substitution rates. These include differential methylation of some transposable elements and potential variation of mismatch repair efficiency.

The factors causing regional CpG substitution rate variation outlined in this paper are important in the study of genome evolution, and in the inference of phylogenetic histories. Our work also highlights the distinct properties of mutations that are dependent on DNA methylation as opposed to those mainly caused by errors during replication.

RESULTS

CpG substitution rate exhibits substantial regional variation

We analyzed approximately 38 million orthologous sites from human, chimpanzee, and baboon obtained by aligning genomic DNA segments from these species. We used only high quality Bacterial Artificial Chromosome (BAC) based sequences in all our analyses. Sequences that resulted from low-coverage whole genome shotgun sequencing projects, which may harbor errors in sequencing and assembly, were not used. We then extracted only non-coding sequences (see Methods) to analyze patterns

of substitution rate variation free from the effect of natural selection. Our final non-coding data set included approximately ~ 14.7 million aligned sites from 17 chromosomes.

We used a parsimony method (DURET 2006; MEUNIER and DURET 2004; MEUNIER *et al.* 2005; TAYLOR *et al.* 2006)to identify sites that have been part of a CpG dinucleotide in the recent past ('CpG sites'). In addition, using the same method, we identified sites that have been part of a GpC dinucleotide ('GpC sites', see the Methods section for details). Because GpC dinucleotides consist of the same bases (C and G) as CpG dinucleotides, while not involved in DNA methylation (RAZIN and RIGGS 1980), they are often used as a dinucleotide control for CpG sites (FRYXELL and MOON 2005; FRYXELL and ZUCKERKANDL 2000; ZHAO and JIANG 2007). We also analyzed all sites that have not been a part of a CpG dinucleotide during the given evolutionary timescale ('non-CpG sites', (MEUNIER and DURET 2004)). Note that GpC sites are a subset of non-CpG sites. Results from non-CpG sites were similar to that from GpC sites (see below).

It is important to note that a great majority of CpG sites in certain regions, called CpG islands, of mammalian genomes are typically unmethylated (BIRD 1986) and hence do not undergo methylation-origin mutation process. Therefore, it is crucial to exclude CpG islands from our analyses. We used similar but slightly more stringent criteria to those proposed by Takai and Jones (TAKAI and JONES 2002), a widely used method, to identify and exclude CpG islands in our data (see Methods).

Table 3.1 describes the numbers of CG->TA transition substitutions in CpG and GpC sites in our dataset. Note that even though there are over an order more GpC sites than CpG sites in our data, the total numbers of CpG and GpC transition substitutions are

Table 3.1. Description of the dataset.

		Intergenic		Introns	
		No. sites	No. substitutions*	No. sites	No. substitutions*
CpG	Repetitive	11257	1761	6941	985
	Non-repetitive	11277	1460	7631	1038
	Total	22534	3221	14572	2023
GpC	Repetitive	154760	1844	84577	827
	Non-repetitive	152695	1619	111815	1106
	Total	307455	3463	196392	1933

***** CG->TA transition substitutions only. The substitutions in humans and chimpanzees were pooled. The number of CpG and GpC sites identified in our dataset and the number of CG-> TA transition substitution in these sites. See *Methods* section for definition of sites. There were 8,971,241 and 5,767,731 aligned sites in intergenic regions and introns, respectively

similar. This observation confirms that CpG substitutions occur much more frequently than other types of substitutions in the human genome (MIKKELSEN T 2005; NACHMAN and CROWELL 2000; SIEPEL and HAUSSLER 2004). For simplicity, we refer to the proportion of the number of transition substitutions to the total number of (CpG or GpC) sites as the 'rate of substitution' in the rest of our paper, which implies that the unit of timescale is since the divergence of the genomes of humans and chimpanzees. The rate of CpG substitution in intergenic regions and introns are 14.29 ± 0.4%, and 13.88 ± 0.5%, respectively. The rate of GpC substitution in intergenic regions and introns are 1.12 ± 0.037% and 0.98 ± 0.043%, respectively.

When all single nucleotide substitutions are considered, their rates vary substantially among different regions, more than expected solely from stochastic effects (MIKKELSEN T 2005). To examine regional variation of CpG substitution rates, we plotted the rate of CpG substitution in 50 kb segments of non-coding regions with at least 10 kb aligned sites (Figure 3.1A). The mean CpG substitution rate in these segments is 15.3%. The observed standard deviation is 6.3%, which is significantly greater than the standard deviation expected under a model that assumes uniform CpG substitution rate in all the segments (95% confidence interval [CI] 4.2% - 5.8%; see Methods). As expected, GpC sites and non-CpG sites also exhibited substantial variation (Figure 3.1B). These results remained the same when we changed the size of the windows examined.

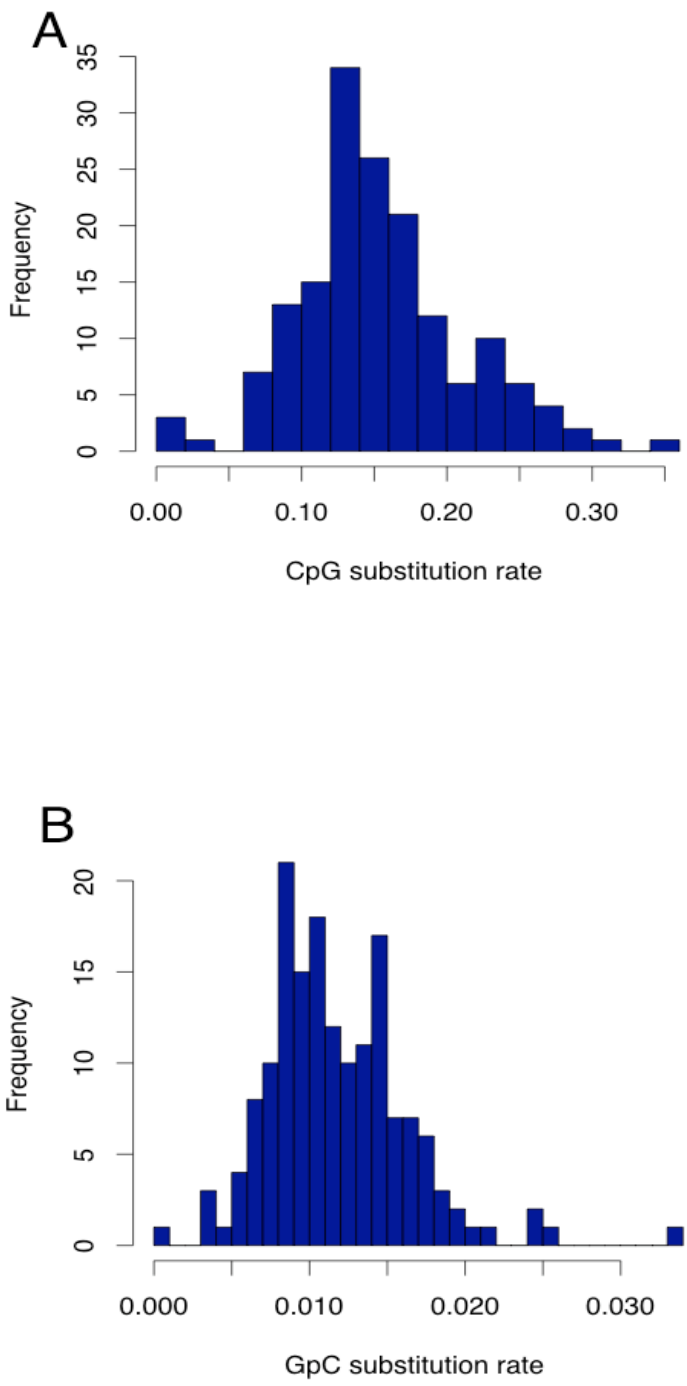


Figure 3.1 Histogram of the rate of CG->TA substitutions in non-coding regions. The rate of CG->TA transitions in CpG sites **(A)**, and GpC sites **(B)** in 50 kb segments of non-coding regions having at least 10,000 aligned sites. Variation of CpG substitution rate among non-coding regions is significantly greater than that expected under a uniform substitution rate model. A similar result was obtained for GpC sites (see text).

The rates of CpG substitution and those of non-CpG substitutions are significantly correlated in our sample (Pearson correlation coefficient $\rho = 0.32$; $P < 0.001$). Because CpG substitution rate did not follow a normal distribution, we also analyzed log-transformed data and obtained a similar result ($\rho_{tr} = 0.31$; $P < 0.001$).

CpG substitution rate in non-coding regions of primate genomes is negatively correlated with G+C content

Earlier studies reported conflicting results on the relationship between the G+C content and the CpG substitution rate of a genomic region. Some studies have proposed a negative relationship (ARNDT *et al.* 2005; FRYXELL and MOON 2005; GAFFNEY and KEIGHTLEY 2005)while others observed no correlation (MEUNIER *et al.* 2005). We analyzed this relationship by dividing the non-coding regions into 6 equal size bins based on their G+C content and plotted the rate of CG->TA substitution at CpG sites in each bin against its average G+C content (Figure 3.2A). To avoid the effect of variation in methylation efficiency, sites in transposable elements were excluded from analyses in this and the following section.

We observed a significant negative correlation between the G+C content of the non-coding regions and the CpG substitution rate (Figure 3.2A; $r^2 = 0.642$, $P = 0.032$). When the CpG data was partitioned into introns and intergenic regions, the negative relationship with G+C content was significant in introns ($r^2 = 0.71$, $P = 0.021$), but not in intergenic regions although there was a clear negative trend ($r^2 = 0.29$, $P = 0.15$). The average length of the intron segments is 5.5kb and of the intergenic segments is 119kb.

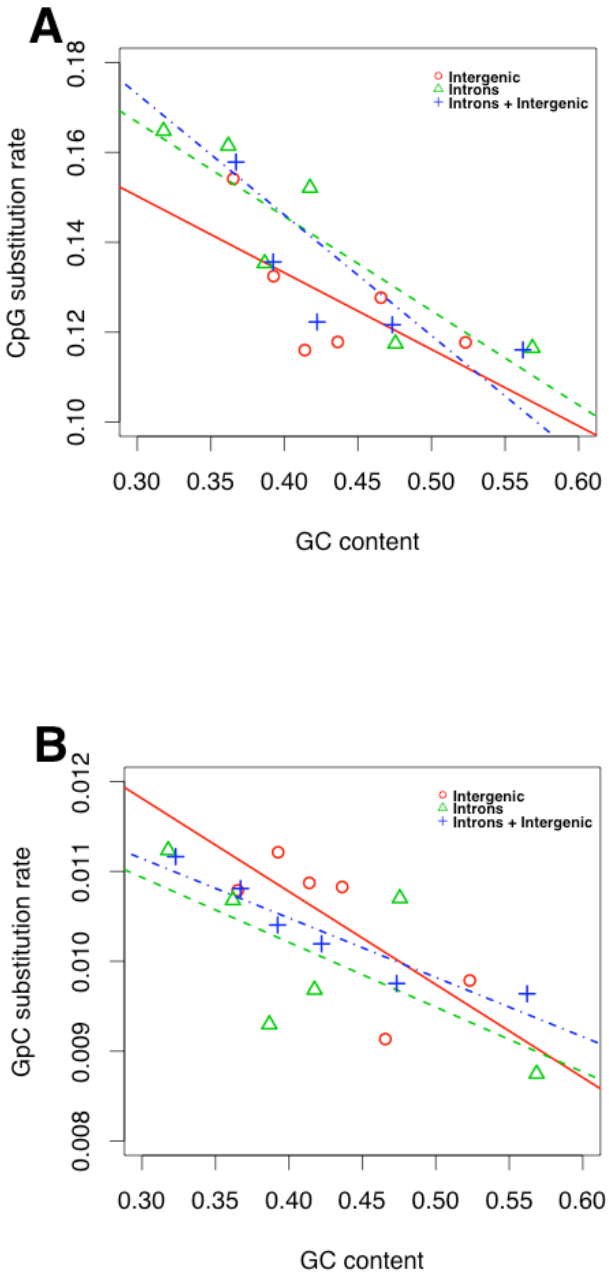


Figure 3.2. Negative correlation between the rate of CG->TA substitution and G+C content of non-coding regions.**(A)** Non-coding regions (intergenic and introns) were partitioned into six equal-sized bins based on their G+C contents. The rates of CG->TA substitutions in CpG sites of these bins are negatively correlated with their G+C contents. The negative relationship holds when introns were analyzed separately. In case of intergenic regions the relationship was not significant. Nevertheless, we found a negative

trend. **(B)** GpC substitution rates in non-coding regions exhibited a negative relationship with G+C contents. When divided into intergenic regions and introns, however, the relationships were not significant, although there was a clear negative trend. Refer text for r^2 values and *P*- values.

We also investigated rates of CG->TA substitutions at GpC sites, as a control for sites that are not affected by DNA methylation (RAZIN and RIGGS 1980). We observed a negative correlation with the G+C content of the non-coding regions (Figure 3.2B; r^2 = 0.86, P = 0.004). When the GpC data was partitioned into introns and intergenic regions, the relationships were not significant [r^2 = 0.313 (P = 0.14), r^2 = 0.341 (P = 0.10) for introns and intergenic regions, respectively]. Nevertheless, we observed negative trends in both introns and intergenic regions (Figure 3. 2B).

Distance-decaying relationship between CpG substitution rate and local G+C content

This negative relationship between CpG substitution rate and G+C content is consistent with the thermodynamic requirement during deamination process (see Introduction). However, given that transition rates at other sites (such as GpC sites) also exhibit negative relationship with G+C content and that thermodynamic hypothesis does not necessarily explain long-range effect of G+C content on CpG substitution rates, we proposed the 'distance-decaying' hypothesis.

We now present our main results on the distance-decaying influence of G+C content on rates of CpG substitution. We performed a sliding window analysis using a window size of 200 bps and a step size of 25 bps (partially overlapping windows) to analyze the relationship between G+C content at varying distances from the CpG site and the rate of CpG substitution. Because the average length of the introns in our dataset was only ~5.5kb [small as compared to the average length of intergenic regions (119kb)], a large proportion of CpG sites in introns may contribute to sliding windows that lie in

exons (the sliding windows extended up to 5kb around each CpG site, see below). Therefore, we used only intergenic CpG sites in this analysis.

At each distance, we binned CpG sites with similar G+C content (as measured from the 200 bps sliding window) at that distance from the site. The cutoffs used for binning were <38%, 38-45%, 45-52%, >52%. These cut-offs divided the data into similar bin size and also roughly corresponded to the traditional definition of isochors (BERNARDI 2000). For each distance *i*, and each G+C content bin, we then calculated the rate of CpG substitution (as the proportion of CpG sites that are mutated). More formally, for each G+C content bin *B*, we considered:

Probability (CpG mutated | window at distance *i* is in G+C content bin *B*).

The results of this analysis are shown in Figure 3A. We observed a clear effect of G+C content of windows close (less than 2000 bps in each direction) to the target CpG site. In particular, higher G+C content in the window lowers the CpG substitution rate (as expected from the negative relationship between G+C content and CpG rate, reported in the previous section). This effect is the most pronounced at distances very close to the target CpG site. For example, ~17% of the CpG sites exhibiting low G+C content (GC < 38%; red color curve in Figure 3. 3A) at distance 100 bps are mutated, while the same measure for CpG sites with high G+C content (GC >52%; black color curve in Figure 3. 3A) at distance 100 bps is ~ 11%. As we move farther away from the CpG site (Figure 3. 3A; left to right along the X axis), the rate of CpG substitution in the low G+C bin (red color curve) and the high G+C bin (black color curve) progressively become closer to each other, displaying the distance decaying effect of G+C content on the rate of CpG substitution.

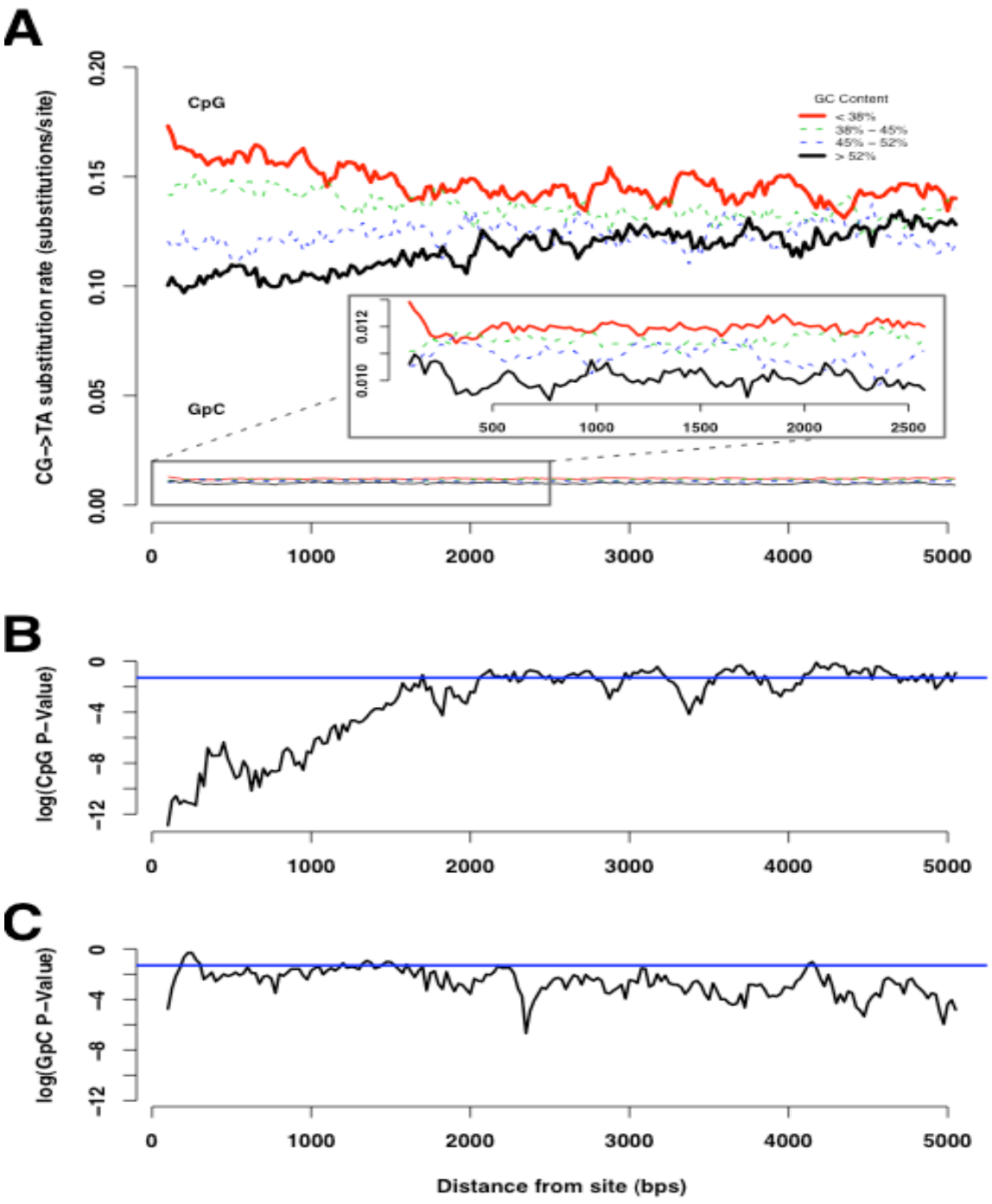


Figure 3.3. Sliding window analysis of the relationship between CpG substitution rate and G+C content of windows. (A): At each distance (along the X-axis), CpG sites were divided into four bins based on G+C content at that distance from the site (as

measured from the G+C content of the 200 bps window centered at that distance; G+C < 38: red curve, 38<=G+C < 45%: green curve, 45 <= G+C < 52%: blue curve, G+C >52%: black curve). The proportion of CpG sites mutated in each of these bins is plotted as a function of distance from the site. At distances closer to the CpG site, the rate of substitution in high local-GC bins (black curve) is clearly lower compared to that of low local-GC bins (red curve). This relationship progressively declines as we move farther away from the site, suggesting a distance decaying relationship between G+C content and CpG substitution rate. In case of GpC sites, we do not observe a distance decaying effect (see inset). **(B):** Results of the chi-square test for the independence of the rate of CpG substitution and the G+C content of the windows at each distance, in log scale. The blue line indicates the *P*-value cutoff of 0.05 [or $\log_{10}$(*P*-value) = -1.30]. The *P*-values are very low at distances close to the CpG site, and progressively becomes larger as the distance from the CpG site increases (distance-decaying effect). The rate of CpG substitution becomes independent of the G+C content [$\log_{10}$(*P*-value) > -1.30] after ~ 2000 bps from the CpG. **(C):** Results of the chi-square test for the independence of the rate of GpC substitution and the G+C content of the windows at each distance, in log scale. Again, the blue line indicates the *P*-value cutoff of 0.05 (or $\log_{10}$(*P*-value) = -1.30). The rate of GpC substitution becomes independent of the G+C content [$\log_{10}$(*P*-value) > -1.30] at a distance very close to the GpC site, and no distance-decaying effect was observed.

This effect appears to vanish at around 2000 bps from the CpG site (i.e., 4000 bps around the CpG site).

We used a chi-square test to determine whether the rate of CpG substitution is dependent on G+C content, at each distance point (Figure 3. 3B). The distance decaying relationship is clear when we observe the *P*-value obtained at each distance. We observe very low *P*-values ($P << 0.05$ or $\log_{10}(P\text{-value}) << -1.30$; which means that the rate of CpG substitution is highly dependent on G+C content) at distances close to the CpG site. As we move farther from the CpG site, the *P*-value increases progressively, becoming insignificant at ~2000 bps, consistent with the results observed in Figure 3A.

We obtained similar results with different overlapping window sizes [for example, window size of 25 bps and step size of 5 bps (Supplementary Fig. B.1)], and when 6 bins were used instead of 4 bins (however, using 6 bins decreased the sample sizes and thus increased the fluctuations in the figures). Similar results were obtained with non-overlapping windows (results not shown). Note that even though there is a distance-decaying relationship between CpG substitution rate and G+C content in immediate neighboring sites, the CpG substitution rate in each GC category differ, although not significantly, from each other even after 2000 bps. In particular, the CpG substitution rate in low GC bins (red curve in Figure 3. 3A) did not converge with the CpG substitution rate in high GC bins (black curve in Figure 3. 3A) even after 2000 bps. In the following section, we show that this may be a consequence of long-range (global) nucleotide composition.

As a control for other dinucleotide sites, we performed the sliding window analysis using GpC sites. In contrast to the case of CpG sites, we do not observe a

distance-decaying relationship between G+C content and the rate of GpC substitution (GpC curves in Figure 3. 3A). The *P*-values for test of independence of substitution rate at GpC sites and G+C content tend to be significant at distances > 2000 bps, suggesting a negative relationship between GpC substitution rate and long-range G+C content (see the following section for more details).

The distance decaying relationship persists after correcting for variation in global G+C content

Although the distance decaying negative relationship between CpG substitution rate and G+C content subsides at ~ 2000 bps (i.e., 4000 bps around the CpG site; Figure 3. 3B), the rate of CpG substitution in low G+C content bin continued to remain higher than that of high G+C content bin (Figure 3. 3A). GpC rates also differed between high and low G+C contents (Figure 3. 3A and 3C).

Indeed, both CpG and GpC rates were significantly negatively correlated with G+C contents of 5kb, 20kb and 100kb blocks around each dinucleotide (results not shown). Also, the G+C contents of 5kb, 20kb and 100kb blocks are all positively correlated (results not shown), presumably because of the isochore structure of primate genomes. In the remainder of this paper, we refer to the G+C content of 100kb segments around each CpG site as GC_{global}, and the negative relationship between GC_{global} and CpG substitution rate as a "*global effect*". We tested if the distance decaying effect of local (<4000 bps) G+C content exists even after controlling for global GC effect (by removing variation in GC_{global}) by the following two analyses.

In the first analysis, we performed the aforementioned sliding window experiment with the G+C content of the windows normalized by the G+C content of the 100kb segment flanking the CpG site. Precisely, we define the normalized G+C content of a window, denoted GC_{norm}, as:

$$GC_{norm} = \frac{GCi}{GC_{global}}$$

where GC*i* denotes the G+C content of the window at distance *i* from the CpG site and GC_{global} denotes the G+C content of the 100 kb region surrounding the CpG site. The value GC_{norm} represents the 'relative' G+C content of the *i*-th window with respect to the G+C content of 100 kb segment within which it belongs.

Figure 3.4 shows the results of the sliding window analysis, where the G+C contents of the 200 bps windows (GC*i*) are normalized. The GC_{norm} cutoffs we used were < 0.9, 0.9 -1.1, 1.1-1.25, > 1.25. These cutoffs were chosen because they divided the data into approximately equal size bins. Even when variation in GC_{global} was removed, we found the same decaying effect of local G+C content on the rate of CpG substitution (CpG curves in Figure 3. 4A, and Figure 3. 4B). The G+C content of windows closer to the CpG site affected the rate of substitution more compared to that of windows farther away from the CpG sites (CpG curves in Figure 3. 4A, and Figure 3. 4B). In contrast to the CpG curves in Figure 3A, the CpG curves in Figure 4A converged, suggesting that the non-convergence in Figure 3A is in fact caused by the global effect. The curves converged at a distance of ~1500 bps, which is slightly lower than that observed in Figure 3A. In the case of GpC sites, as expected, there was no distance decaying relationship and

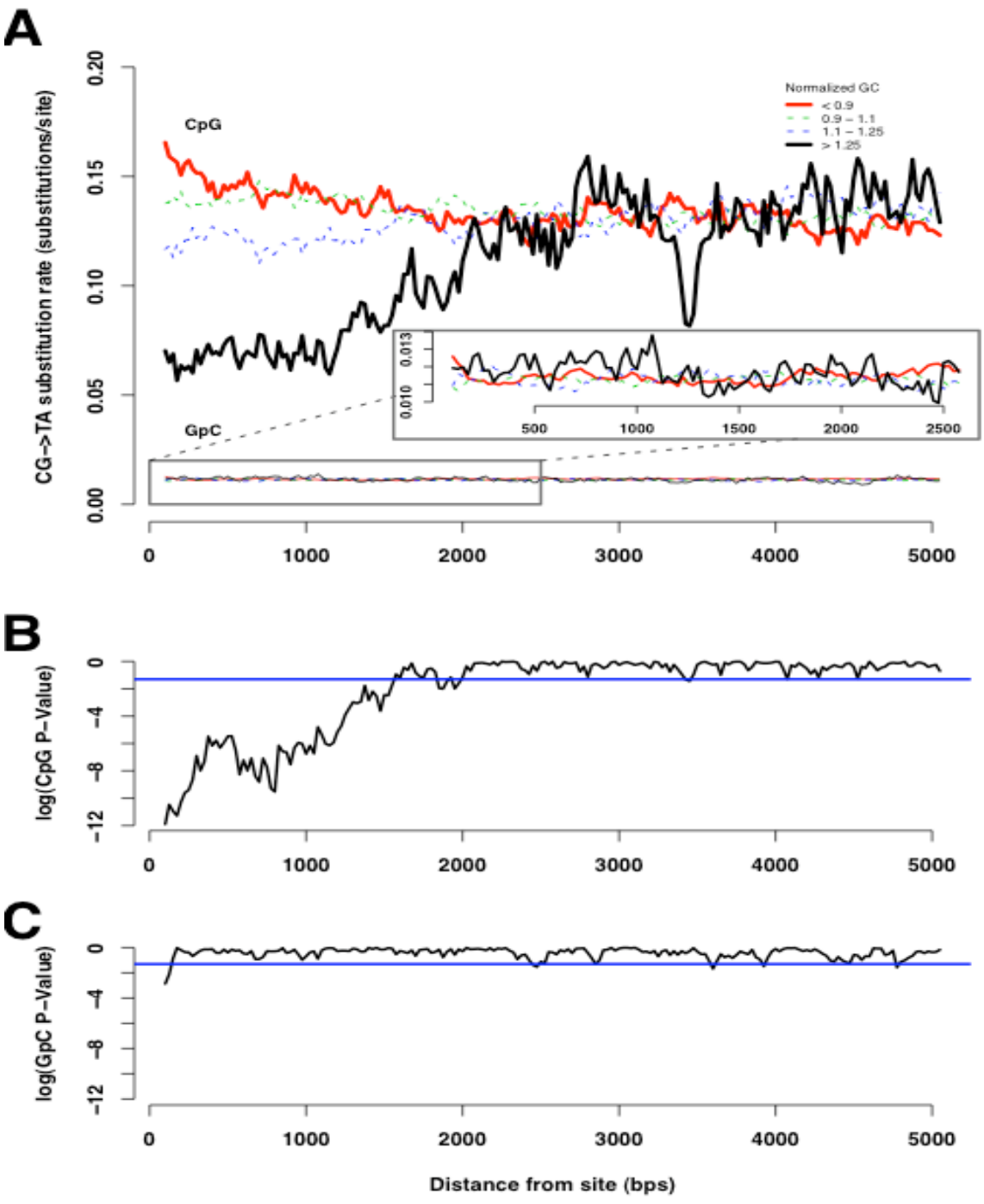


Figure 3.4 Sliding window analysis of the relationship between CpG substitution rate and normalized G+C content. The same experiment as in Figure 3 with the G+C content of each window normalized with respect to GC_{global} (removing global effect). **(A):** The distance decaying effect of G+C content on the rate of CpG substitution persists even

after removing the global effect. In case of GpC substitutions, there was no distance decaying effect. **(B):** Results of the chi-square test for the independence of the rate of CpG substitution and the G+C content of the windows. The blue line indicates $\log_{10}$ (*P*-value) = -1.30. The distance decaying effect subsided after ~1500 bps **(C):** Results of the same experiment as in panel B, but for GpC sites. There is no distance decaying effect, as expected.

the curves converged at a distance very close to the site (GpC curves in Figure 3. 4A and 4C).

In our second analysis, we first focused on the distribution of GC_{global}. GC_{global} exhibits a bimodal distribution with means of ~39% and ~48%, respectively (Supplementary Fig. B.2A). A similar distribution of G+C content was observed in case of 100kb segments surrounding GpC sites (Supplementary Fig. B.2B). These bimodal distributions occur because the number of CpG sites and GpC sites are expected to increase with the G+C content of the non-coding region. To reduce variation in GC_{global}, we analyzed these two distributions (which we call as low- GC_{global} and high- GC_{global} regions) separately.

We performed the sliding window analysis to test for the distance decaying effect of local G+C content in low- GC_{global} and high- GC_{global} regions (without normalizing). In the case of CpG sites in low- GC_{global} regions (Supplementary Fig. B.3A, and B.3B), we obtained similar results as in Figure 4A. In the case of CpG sites in high- GC_{global} regions, there were large fluctuations in the trend, especially in the low local-GC bins (Supplementary Fig. B.4A; red CpG curve). These fluctuations are caused by the reduced sample size in this bin (data not shown). Nevertheless, the distance decaying effect is clear when we consider the curves of high- and low local-G+C content bins (black and red curves CpG, respectively, in Supplementary Fig. B.4A). Consistent with the results from Figure 4, the distance from the CpG site at which the curves converged was ~1500 bps (Supplementary Fig. B.3A). Again, GpC sites did not exhibit any distance decaying effect (Supplementary Figs. B.3A, B.3C and B.4A, B.4C).

As a complementary analysis, we performed the sliding window analysis on CpG and GpC sites that are randomly sampled from low- GC_{global} and high- GC_{global} regions and obtained similar results (Supplementary Fig. B.5).

These results suggest that that even after removing the global effect, the distance-decaying effect of local G+C content on CpG substitution rate persists, and it subsides after approximately 1500 bps in each direction from each CpG site. These results were robust to using the flanking 5kb or 20kb window for the normalization, different window sizes, and non-overlapping windows (results not shown).

Other factors that affect CpG substitution rates

The previous sections analyzed the effect of the efficiency of deamination step on CpG substitution rates. In this section we briefly discuss other factors that cause regional heterogeneity of CpG substitution rates.

For this purpose we consider that a CpG substitution occurs in a three-step process (this is a simplification, but for our purposes this view suffices). First, DNA methylation in mammals (and other warm-blooded vertebrates) occurs specifically at the cytosine bases of CpG dinucleotides. Second, the methylated cytosine undergoes spontaneous deamination. Deaminated methyl-cytosine is identical to thymine, creating a C to T mismatch at the CpG dinucleotide. Third, if this mismatch is left un-repaired, it causes a C to T transition (or a G to A transition in the complementary strand) in the next replication cycle. Based upon such molecular mechanism of the origin of CpG substitution, their rates may vary due to differences in (i) levels of germline methylation,

(ii) efficiency of deamination of a methylated cytosine, and (iii) repair efficiency of a transition that occurred via deamination of a methylated cytosine.

A strong support for the first cause, differential methylation, is the observation that transposable elements have generally higher rates of CpG substitutions (MEUNIER *et al.* 2005). Proliferation of transposable elements can potentially impose several deleterious effects on a genome, because insertion of transposable elements can disrupt gene regulation, or cause deleterious recombination events, translocations and other rearrangements (YODER *et al.* 1997). Hence, some organisms may use DNA methylation as a defense mechanism against the proliferation of transposable elements (MEUNIER *et al.* 2005; YODER *et al.* 1997).

In our data, we found that transposable elements as a whole exhibit ~14% higher CpG substitution rate as compared to that of non-repetitive regions (see Supplementary Text in Appendix B). Overall, there is a significant positive correlation between CpG rates and the proportion of transposable elements in a non-coding segment ($\rho = 0.142$; $P < 0.001$ and $\rho_{tr} = 0.199$; $P < 0.001$), consistent with the idea that transposable elements have higher CpG substitution rates as a consequence of increased methylation.

Interestingly, when transposable elements were divided into different classes of elements (LTR, LINE, SINE and DNA elements), SINEs showed similar level of CpG rates to non-repetitive sequences. This was also observed by (MEUNIER *et al.* 2005), who proposed that it might be a consequence of the complex methylation pattern of SINEs in germline cells (CHESNOKOV and SCHMID 1995). However, another possibility is that SINEs show lower CpG rate because their G+C content is in general higher compared to other classes of transposable elements, in light of our finding that there is a negative

relationship between CpG substitution rates and G+C contents (Results above). SINEs have the highest G+C content among the classes of transposable elements considered in our analyses (50.7% G+C content while 36.6%, 44.0%, 36.4%, for LINEs, LTRs and DNA transposons). This mechanism can at least partially explain why SINEs have lower rates of CpG substitutions. Because a major proportion of CpG sites in the transposable elements of introns were contributed by SINEs, the rate of CpG substitution in intronic transposable elements is comparable to that of intronic non-repetitive regions (see Supplementary Text in Appendix B).

If the efficiency of the third step, namely mismatch repair, affects regional variation of CpG substitution rates, we expect to see a significant negative correlation between recombination rates and CpG substitution rates. Indeed, we did find a significant negative correlation between recombination rates and the rates of CpG transition substitution (ρ = -0.165; $P < 0.001$). This can be caused by the direct effect of biased gene conversion acting on mutations at CpG sites, repairing the TG mismatches (caused by deamination of methyl-cytosine) to correct CG basepairs, the efficiency of which depends on the rate of recombination. Alternatively, biased gene conversion may increase G+C content, which in turn reduces the intensity of DNA melting, leading to a lower rate of substitution at CpG sites. In the first scenario, a significant correlation between CpG substitution rate and recombination rate is expected independent of the effect of G+C content. In the second scenario, the effect of recombination may disappear after correcting for G+C content. In our case, there was no significant correlation after removing the effect of G+C content using partial correlation analysis (partial correlation coefficient = -0.03; $P > 0.05$). However, this question needs to be revisited with more

data, most certainly when a higher-quality whole genome alignment of human, chimpanzee, and rhesus macaque becomes available.

DISCUSSION

Even though the prevalence and importance of CpG substitutions are long recognized (BIRD 1980), the causes and patterns of their regional variation have been relatively little explored. In this study, we analyzed a large amount of high-quality genome sequence data (based upon BAC-clones) to provide a better understanding on the patterns and causes of variation of CpG substitution rates in primate genomes.

The negative relationship between CpG substitution rate and G+C content is consistent with the idea that variations in the efficiency of the deamination process (FREDERICO *et al.* 1993; FRYXELL and MOON 2005; FRYXELL and ZUCKERKANDL 2000) cause regional variation of CpG substitution rates. On the other hand, biased gene conversion can also explain this relationship. Even though we cannot distinguish the relative contributions of these two processes by our analysis, the distance decaying relationship between G+C content and CpG substitution rate strengthens the former hypothesis.

Our distance-decaying hypothesis is contradictory to conclusions drawn by Fryxell and Moon (FRYXELL and MOON 2005) and Zhao and Jiang (ZHAO and JIANG 2007). Fryxell and Moon conclude that G+C content of short and long segments surrounding each CpG site has similar degrees of correlations with the rate of CpG substitutions; this is clearly counter to the distance decaying influence hypothesis. However, their analysis was based on a considerably smaller dataset (a total of 4437 CpG

and GpC sites, as compared to 497467 sites used in this study; see Table 3.1), and only two length scales (564 bps and 163kb) were analyzed. In another study, Zhao and Jiang analyzed a larger human SNP dataset (292216 CpG and GpC sites) and observed that the absolute value of the slope of the relationship between log (CpG substitution rate) and G+C content *increased* with the length over which G+C content was measured (segment lengths of 101 bps to 1001 bps around the CpG site). Their results are counter to the thermodynamic mechanisms which were first suggested by Fryxell and Zuckerkandl (FRYXELL and ZUCKERKANDL 2000) and now drive our hypothesis. The discrepancy is likely due to the different method and dataset used by Zhao and Jiang, as compared to our study (See supplementary material in Appendix B for details). Our results in this vein support the thermodynamic mechanism.

Our analyses (Figs. 3 and 4) show that the CpG substitution rate has a significantly stronger relationship with the *local* G+C content. The influence of the G+C content decays with the distance to the target CpG site, subsiding at around 1500-2000 bps. There are at least two immediate questions raised by our analysis. First, what are the causes of the local distance decaying relationship, which extends up to 1500-2000 bps. Given the absence of evidence for the effect of transcription coupled repair or transcription induced deamination (only intergenic regions were analyzed), or change in mutational biases over time (ARNDT *et al.* 2003), we propose that the distance decaying relationship may be a consequence of the poorly understood mechanism of DNA melting (LILLEY 1988), which is required for the deamination process.

The second question is whether there is both global and local effect of G+C content on CpG substitution rate. We raise this question because it is possible that only

the local G+C content may affect the CpG substitution rate, and the global effect is a consequence of the so-called 'isochore' structure, which causes a positive correlation between local G+C content and global G+C content (BERNARDI 2000; EYRE-WALKER and HURST 2001; NEKRUTENKO and LI 2000). Although the answer to this question is still unclear and more direct experiments are required, the convergence of CpG curves when the data is divided into low-GC_{global} and high-GC_{global} distributions (For example, Supplementary Figs. B.2A, B.2B) despite the fact that there is some variation in GC_{global} even within these distributions (Supplementary Fig. B.1A) suggests that GC_{global} may not have a direct effect on the rate of CpG substitution.

We found a negative relationship between GpC substitution rate and G+C content (Figure 3. 2B), in agreement with the result obtained by a previous study (FRYXELL and MOON 2005). As suggested by Fryxell and Moon (FRYXELL and MOON 2005), this relationship can be explained by biased gene conversion (GALTIER *et al.* 2001) or by regional differences in the deamination of unmethylated cytosines (an underlying mutational bias). In the absence of a distance decaying relationship between GpC substitution rate and local G+C content (Figs. 3A and 4A), which is expected if deamination plays a major role in causing GpC substitution rate variation, it seems more likely that biased gene conversion is the cause of the negative relationship between G+C content and GpC substitution rate.

Our results demonstrate the significance of the distinctive properties exhibited by methylation dependent substitutions versus substitutions caused by other molecular mechanisms. Models of molecular evolution and methods of phylogenetic inference, especially those concerned with mammalian and bird genomes (which have high rates of

CpG substitutions), may benefit by considering the local effect of G+C content on CpG substitution rates (e.g., (ARNDT *et al.* 2005; HWANG and GREEN 2004; SIEPEL and HAUSSLER 2004)).

METHODS

General approach

We can use a parsimony method (MEUNIER and DURET 2004; MEUNIER *et al.* 2005; TAYLOR *et al.* 2006) to infer CpG sites and determine the substitution rate in humans and chimpanzees using baboon as an outgroup. However, we cannot distinguish substitutions caused by methylation followed by deamination versus replication errors. In primates, CpG substitutions are markedly more frequent than other single nucleotide substitutions (MIKKELSEN T 2005; NACHMAN and CROWELL 2000). Therefore we assumed methylation followed by deamination is the primary cause of all CpG substitutions in these genomes.

Human, chimpanzee and baboon sequences

We analyzed ~38 Mbps of sequence data obtained from two different sources; 22 Mbps from sequences orthologous to the human chromosome 7 sequenced by the NISC comparative sequencing group, and ~15 Mbps obtained from other chromosomes from database mining.

Sequence data from chromosome 7: We analyzed ~22 Mbps from human chromosome 7. The sequences are the same as that in Dataset 1 of (ELANGO *et al.* 2006); however, we made our analysis more stringent by removing introns that are alternatively spliced (see

below). Chimpanzee and baboon BAC clones orthologous to the regions in human chromosome 7 were isolated and sequenced as described in (THOMAS *et al.* 2002).

Sequence data from other chromosomes: For other chromosomes, we mined sequence data from GenBank (BENSON *et al.* 2006). Briefly, we downloaded all the baboon and chimpanzee BAC clones available as of March 2006 from GenBank and then used Blastz (SCHWARTZ *et al.* 2003) to determine the high scoring matching segments. Then, we used a pipeline of programs to perform chaining and netting, as described in (KENT *et al.* 2003), to establish orthology. A detailed description of the procedure used is presented in the supplementary material (see Apendix B).

Sequence annotation and alignment

Sequences orthologous to human chromosome 7 were aligned using the Threaded Blockset Aligner program (BLANCHETTE *et al.* 2004). The "best chains" (See Appendix B) from other chromosomes were aligned using the Multiz program (BLANCHETTE *et al.* 2004). Non-coding regions (introns and intergenic regions) were identified using gene annotations included in the Known Genes and Ensembl Genes tables of the *hg*17 assembly of the human genome at UCSC Genome Browser (KENT *et al.* 2002). Intergenic and intronic sequences likely to be selectively constrained [the 5' and 3' untranslated regions, first introns and small (<250 bps) introns or intergenic intervals] were excluded. In addition, we excluded alternatively spliced introns based on gene annotations from the UCSC genome browser. In particular, if the span of an intron is different in different transcripts of the same gene, it was removed. The above methods yielded ~14.7 Mbps of aligned sites in the non-coding regions of the human genome (Table 3.1). Recombination rates were obtained from (MYERS *et al.* 2005).

Identification of CpG islands

CpG islands are regions of the genome where a majority of CpG sites are not methylated. Because methylation of the cytosine in CpG dinucleotides is a prerequisite for the CG->TA mutation to occur (see Introduction), it is crucial for our analyses to exclude CpG sites in CpG islands. Takai and Jones (TAKAI and JONES 2002) proposed that a good definition of CpG islands is- a region of the genome with (a) G+C content > 55%, (b) length > 500 bps and (c) observed/expected proportion of CpG dinucleotides (OE) > 0.65. The condition length > 500 bps was used by Takai and Jones to eliminate the possibility of falsely calling a CpG rich regions generally associated with Alu elements as CpG islands. However, there may be some CpG islands less that 500 bps in length.

In this study, we changed the length constraint to > 200 bps to err on the side of caution to safely eliminate most of the CpG islands. We masked out CpG islands identified using the algorithm by Takai and Jones (TAKAI and JONES 2002) with parameters G+C content > 55%, OE > 0.65, length > 200.

Identification of CpG sites

To identify CpG sites, we used a parsimony method. Specifically, CpG sites are the middle base of the sites having the following human/chimpanzee/baboon patterns: XNG/XCG/XCG or XCG/XNG/XCG, where X is any nucleotide except G. Given the evolutionary distances considered here, in spite of the hypermutability of CpG sites, such a definition is shown to be quite accurate by a simulation study in Meunier and Duret (MEUNIER and DURET 2004). GpC sites are sites having the following

human/chimpanzee/baboon pattern GNY/GCY/GCY or GCY/GNY/GCY, where Y is any nucleotide except G. The restrictions on X and Y were imposed to avoid overlapping CpG and GpC sites. Non-CpG sites are defined as sites not preceded by a C and not followed by a G. Sites following the complementary patterns of the above definitions were also considered as CpG and GpC and non-CpG sites.

Substitution rate estimates and statistical tests

To obtain better estimates of substitution rates, we pooled together substitutions in human and chimpanzee lineages. The rate of substitutions for a particular class (CpG or GpC) of site was estimated by dividing the number of substitutions by the total number of sites in that class.

Simulation of uniform rate model

To test if the observed variation in CpG substitution rate in the 50 kb segments is greater than that expected under a uniform substitution rate model, we performed the following simulation. For each segment (s), we kept the number of CpG sites (n_s) the same as that observed in our data and sampled the number of substitutions from the binomial distribution $b(n_s,p)$, where p is the probability of observing a CpG substitution. p was kept constant across segments (0.153 in the case of CpG sites). We simulated 1000 replicates and found the standard deviation of CpG substitution rate among the segments in each replicate. The range between 2.5% quantile and 97.5% quantile was taken as the 95% confidence interval of the standard deviation under uniform substitution rate model. A similar simulation was also performed for GpC and non-CpG sites.

ACKNOWLEDGEMENTS

S.V.Y is supported by funds from the Georgia Institute of Technology. E.V is supported by NSF career grant. This research was supported in part by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health. Suggestions from Adam Eyre-Walker regarding the distance-decaying effects are greatly acknowledged. We are thankful to James W. Thomas and Phil Green for helpful comments on a previous version of this manuscript.

CHAPTER 4

DNA METHYLATION AND STRUCTURAL AND FUNTIONAL BIMODALITY OF VERTEBRATE PROMOTERS

ABSTRACT

Human promoters divide into two classes, the low CpG (LCG) and the high CpG (HCG), based on their CpG dinucleotide content. The LCG class of promoters is hyper-methylated and is associated with tissue-specific genes, while the HCG class is hypo-methylated and associated with broadly-expressed genes. By analyzing several chordate genomes separated for hundreds of millions of years, here we show that the divide between low CpG and high CpG promoters is conserved in several distantly related vertebrate taxa (including human, chicken, frog, lizard, and fish), but not in close invertebrate outgroups (sea squirts). Furthermore, LCG and HCG promoters are distinctively associated with tissue-specific and broadly expressed genes in these distantly related vertebrate taxa. Our results indicate that the function of DNA methylation on gene expression is conserved across these vertebrate taxa and suggest that the two classes of promoters have evolved early in vertebrate evolution, as a consequence of the advent of global DNA methylation.

INTRODUCTION

Analyses of human and mouse promoters have revealed an intriguing structural and functional bimodality (CARNINCI 2006; SAXONOV *et al.* 2006; TANG and EPSTEIN 2007; WEBER *et al.* 2007). Structurally, they divide into two distinct classes based on

CpG dinucleotide content – the low CpG content (LCG) and the high CpG content (HCG) promoters; the latter being associated with CpG islands (SAXONOV *et al.* 2006). At the functional level, LCG-genes tend to be tissue- specific while HCG-genes are broadly expressed. A common explanation for this phenomenon is germline DNA methylation (ANTEQUERA 2003; SAXONOV *et al.* 2006; WEBER *et al.* 2007), because methylation influences both sequence property and gene expression.

In terms of its effect on sequence, DNA methylation is highly mutagenic. In vertebrate genomes, methylation occurs almost exclusively at the cytosines in CpG dinucleotides. Methylated cytosines undergo rapid deamination to become a thymine, causing a C-to-T transition (or G-to-A on the complementary strand) (COULONDRE *et al.* 1978). In the human and chimpanzee genomes, for example, methylation-origin transitions occur almost 15- fold more frequently than other single nucleotide substitutions (see Chapter 2 and 3). As a consequence, CpG dinucleotides are depleted from methylated regions over evolutionary timescale (BIRD 1980). Thus, normalized CpG content (CpG O/E or 'CpG content' in the rest of the paper; See Methods) is an indicator of the level of DNA methylation (low CpG content and high CpG content reflect hyper-methylation and hypo-methylation, respectively: (SUZUKI *et al.* 2007; WEBER *et al.* 2007)).

In terms of function, at least in mammals, promoter methylation can dampen gene expression. This is achieved either directly by interfering with transcription factor binding, or indirectly through recruitment of methyl-CpG-binding proteins to alter chromatin structure (JONES and TAKAI 2001; KLOSE and BIRD 2006). LCG promoters may therefore be associated with (somatic) tissue-specific genes as a consequence of

germline DNA methylation, while the promoters of broadly expressed genes remain unmethylated thereby forming the HCG class (VINOGRADOV 2005).

While the role of mammalian promoter methylation and its effect on CpG content and expression breadth is well studied (ANTEQUERA 2003; CARNINCI 2006; SAXONOV *et al.* 2006; TANG and EPSTEIN 2007; WEBER *et al.* 2007), the origin and evolution of such phenomenon is little understood. Here we investigated these aspects, by first focusing on the observation that the patterns of DNA methylation differ greatly among diverse taxa. Mammalian genomes exhibit a global DNA methylation pattern (~80% of the CpGs are methylated) in most cell types ((TWEEDIE *et al.* 1997) and the references therein), and thus are largely depleted of CpG dinucleotides (BIRD 1980). Most of the vertebrate genomes analyzed are similarly globally methylated (BIRD 1980; TWEEDIE *et al.* 1997).

Studies indicate that such 'global' genomic methylation is restricted to vertebrates. The Invertebrate animals distantly related to vertebrates such as *Drosophila* and *C. elegans* generally lack germline DNA methylation (TWEEDIE *et al.* 1997). Genomes of close outgroup of vertebrates, such as genomes of invertebrates within the chordate phylum (e.g., urochordate sea squirt and cephalochordate amphixious), and echinoderms (e.g. sea urchin) exhibit a "mosaic" CpG methylation pattern with long methylated regions and equally long unmethylated regions. Based upon these, it is proposed that the transition from mosaic to global methylation pattern have occurred early in vertebrate evolution (TWEEDIE *et al.* 1997).

Interestingly, the aforementioned functional role of promoter DNA methylation may also be unique to vertebrates. A recent study (SUZUKI *et al.* 2007) showed that

methylation at CpG sites in the urochordate *Ciona intestinalis,* which exhibits a mosaic methylation pattern, is targeted to the intragenic regions of a subset of genes. Thus, the intragenic regions (in contrast to promoters in humans) fall into low-CpG and high-CpG categories in this genome. Promoters analyzed in (SUZUKI *et al.* 2007) were not preferentially targeted by DNA methylation.

Several questions emerge when we synthesize these observations: does the structural bimodality in mammalian promoters exist in other vertebrates? If yes, does the relationship between structural bimodality and expression breadth hold in those species? When did the promoter bimodality evolve? If the structural bimodality coincides with the functional bimodality in distantly related vertebrate species, we can infer that DNA methylation is the underlying link for both phenomena, and that structural and functional promoter bimodality has evolved early in vertebrate evolution, rather than independently several times in different vertebrate taxa. In this study we provide answers to some of these questions, and propose a model for the evolution bimodal vertebrate promoters.

MATERIALS AND METHODS

Genome sequences and annotations

We analyzed six vertebrate and three invertebrate genomes, covering substantial phylogenetic depth. The genome build, source, and gene annotations used in the study are shown in Supplementary Table C.1. We present results from the analyses of the following genomes in the main text: zebrafish (*Danio rerio*), frog (*Xenopus tropicalis*), chicken (*Gallus gallus*), human (*Homo sapiens*), and a sea-squirt (*Ciona intestinalis*), because these genomes had relatively large numbers of curated RefSeq (PRUITT *et al.*

2007) gene annotation data (Supplementary Table C.1). For the human genome, we verified our results by using experimentally characterized highly accurate Transcription Start Site (TSS) annotations available in the Database of Transcription Start Sites (DBTSS; (WAKAGURI *et al.* 2008).

In all the analyses we removed genes with more than one transcript, to avoid errors in TSS annotation. Promoters were defined as 600 bps regions upstream of TSS. Qualitative results of our analyses did not change when we used 1kb upstream regions as promoters. We also removed promoters that lied within a distance of 3kb from any other gene.

Because natural selection on coding sequence could potentially confound our results, exons were removed from all the intragenic analyses. First introns were also removed from our analyses, because they may encode regulatory elements (MAJEWSKI and OTT 2002).

Non-first introns may still harbor some regulatory sequences; for example, some introns may carry CpG islands (GARDINER-GARDEN and FROMMER 1987). The presence of intronic CpG islands may have caused the skew toward greater CpG O/E in human and chicken introns (Figure 4.1I, J).

Repetitive elements were identified using repeat masker annotations in the UCSC genome browser (KENT *et al.* 2002). We masked repetitive elements from our analyses, because CpG contents in repetitive elements does not faithfully represent the historical methylation status of the region due to differences in the time of insertions. However, including repetitive elements in the analyses did not change our qualitative results (results not shown). We note that in case of zebrafish, including repetitive elements has an effect

of increasing overall CpG O/E, which is due to the recent origin of repetitive elements in this genome (N. Elango and S. Yi, unpublished results).

Measurement of normalized CpG contents

The 'normalized CpG content' (CpG O/E) is defined as

$$CpG[O/E] = \frac{P_{CpG}}{P_C * P_G}$$

where $\boldsymbol{P}_{CpG}$, $\boldsymbol{P}_c$ and $\boldsymbol{P}_G$ are the frequencies of CpG dinucleotides, C nucleotides, and G nucleotides, respectively. Several studies (e.g.,(SUZUKI *et al.* 2007; WEBER *et al.* 2007)) confirmed experimentally that CpG O/E is a reliable measure of methylation status.

Statistical test for bimodal distribution

The unimodality or bimodality of normalized CpG content distributions was tested using the NOCOM software. Briefly, the software uses an expectation maximization algorithm to fit the data to both unimodal and bimodal distribution models, and finds the maximum likelihood values (L_0 and L_1 for unimodal and bimodal models, respectively). To test if the bimodal distribution model is a better fit to the data as compared to the unimodal distribution model, a statistic $G^2 = 2\ [\ln(L_1) - \ln(L_0)]$ was calculated. This statistic approximately follows a Chi-square distribution with two degrees of freedom.

Analysis of Expression data

Expression data were obtained from EST counts in the Unigene database (WHEELER *et al.* 2008). Genes with EST count ≥ 1 in a tissue were considered to be expressed in that tissue. The expression breadth of a gene is the number of tissues in

which it is expressed. For the human genome, we additionally analyzed two microarray datasets. The first dataset contains expression data from 79 tissues measured using 3' arrays (SU *et al.* 2004). Genes with average difference value >200 in a tissue were considered expressed in that tissue (SAXONOV *et al.* 2006; SU *et al.* 2004). The second dataset contains exon array expression data from six tissues (XING *et al.* 2007), namely heart, kidney, liver, muscle, spleen, and testis. The probes in the exon array are evenly spaced and dense [147 probes per gene, compared to 11 probes per gene in the 3' array (XING *et al.* 2007)]. Therefore exon arrays are considered to provide more accurate measures of gene expression compared to the 3' arrays used in (SU *et al.* 2004) (KAPUR *et al.* 2007; XING *et al.* 2007).

RESULTS

Patterns of intronic and promoter methylation in invertebrate genomes closely related to vertebrates

We analyzed patterns of CpG dinucleotide depletion in upstream promoter regions and intragenic regions of several invertebrate and vertebrate genomes, and related them to the known patterns of genomic methylation. Previous studies have shown that while most vertebrate genomes are globally methylated in many different tissue types ((TWEEDIE *et al.* 1997) and references therein), invertebrates closely related to vertebrates such as urochordate (e.g., sea squirt) and echinoderms (e.g., sea urchin) exhibit 'mosaic' pattern of genomic methylation (SIMMEN *et al.* 1999; SUZUKI *et al.* 2007; TWEEDIE *et al.* 1997). In particular, a recent study demonstrated that in the genome of a sea squirt (*C.*

intestinalis), methylation is targeted to only a subset of intragenic regions (SIMMEN *et al.* 1999; SUZUKI *et al.* 2007; TWEEDIE *et al.* 1997).

We compared normalized CpG content (CpG O/E) of upstream promoter regions (defined as 600 bps upstream of the transcription start site (TSS): results remained the same when 1kb instead of 600bps of upstream regions were analyzed) and of introns. Because our purpose is to compare patterns of intragenic methylation versus upstream promoter regions, we did not include promoter regions downstream of TSS, which often includes first exons and introns. Intron CpG O/E serves as an indicator of the level of genome-wide methylation. Intergenic CpG O/E is not used for this purpose because these genomes differ greatly in terms of genome size and the amount of intergenic regions. In the relatively well-annotated human genome, for example, CpG O/E distributions of intergenic and intronic regions are similar (Supplementary text and Figure C.1 in Appendix C).

Introns of *C. intestinalis* show two distinctive distributions, one with the mean CpG O/E ~1 and the other with the mean CpG O/E around 0.5 ((SUZUKI *et al.* 2007); Figure 4.1F). This observation is in accord with the finding that the genomic methylation pattern is 'mosaic' in *Ciona* genome, where intragenic regions of a subset of genes are methylated and others are not methylated (SUZUKI *et al.* 2007). The introns with high CpG O/E (~1) represents non-methylated genes, and those with low CpG O/E (~0.5) showcases methylated genes ((SUZUKI *et al.* 2007); Figure 4.1F). In contrast, we found that CpG content of *Ciona* promoters follows a unimodal distribution with its mean ~1 (CpG occurs at the expected frequency: Figure 4.1A) indicating that promoters are largely unmethylated in this genome.

We analyzed distributions of CpG O/E in promoters and introns of another sea squirt, *C. savignyi*. Genetic divergence between *C. intestinalis* and *C. savignyi* is known to be similar to that between human and chicken (SMALL *et al.* 2007). We obtained similar results (Figure S2). The promoter regions follow a unimodal Gaussian distribution with mean ~1, while the intragenic regions show two distinctive curves.

We also analyzed data from sea urchin (*Strongylocentrotus purpuratus*), which also exhibits a patch DNA methylation pattern (TWEEDIE *et al.* 1997). There are only 131 RefSeq genes from this species that satisfy the criterion of single transcript. Even with this small sample size, results from the promoter and intronic regions of this species are similar to those in the two *Ciona* genomes (Figure S2).

This above pattern in invertebrate genomes is opposite to that in the human genome. As discussed earlier, human promoters exhibit bimodality of hypo- and hyper-methylated portions (Figure 4.1E) whereas introns show a unimodal distribution with mean around 0.2, reflecting heavy global methylation (Figure 4.1J).

Distribution of promoter CpG content is bimodal in distantly related vertebrate species.

Next we investigated if the pattern of promoter and intron CpG depletion found in humans is conserved in non-mammalian vertebrates. We analyzed the following distantly related vertebrate genomes, in addition to the human genome: fugu (*Takifugu rubripes*), zebrafish (*Danio rerio*), frog (*Xenopus tropicalis*), lizard (*Anolis carolinensis*), and chicken (*Gallus gallus*). Among these species, results from fugu and lizard are presented in the supplementary material (Figure C.2) because of the potential inaccuracy

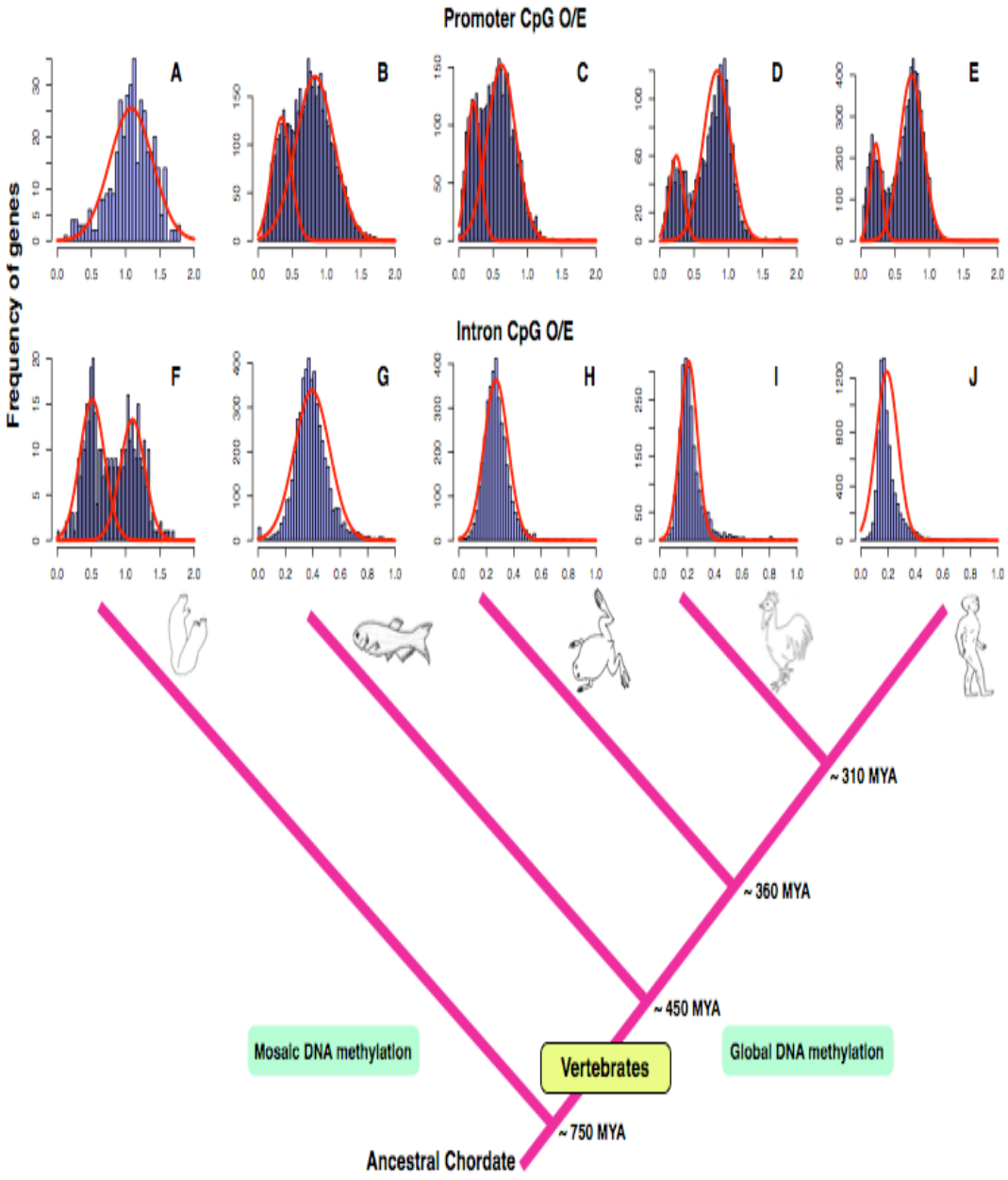


Figure 4.1. Contrasting distributions of normalized CpG contents (CpG O/E) of vertebrate and invertebrate promoters and introns. The distributions of normalized CpG contents (CpG O/E) in 600 bps region upstream of protein coding genes (A-E), and introns (F-K) of studied genomes. Approximate timescale and evolutionary relationships among the studied genomes are shown below the distributions (Hedges and Kumar 2003). The best-fitting normal distributions are shown above the observed distributions.

of annotations. We will focus on results from the four remaining vertebrate genomes, which are relatively well annotated and cover sufficient phylogenetic depth (Figure 4.1).

Intronic CpG content of the four vertebrate genomes shows clear unimodal distributions (Figure 4.1F-J). The mean intronic CpG O/E are all below 1, indicating that these genomes are heavily methylated. The suppression of CpG frequency is most pronounced in the human and chicken introns, where the mean CpG O/E is 0.23 and 0.22, respectively (Figure 4.1I,J; Table 4.1). In the zebrafish, CpG frequency is approximately 40% of the expected (Figure 4.1G). The observed inter-species differences in intronic CpG O/E could be due to differential methylation levels, or due to the variability of the efficiency of deamination, which is a key step in methylation-induced CpG mutations (FREDERICO *et al.* 1993).

In contrast, upstream promoter regions of all four vertebrate genomes follow two distinctive distributions of LCGs and HCGs (Figure 4.1 and Table 4.1). We used an expectation-maximization (EM) algorithm to fit the observed distributions to unimodal, bimodal and trimodal Gaussian distributions and compared the likelihoods (Materials and Methods, Table 4.1). Bimodality is consistently a far better fit to the observed distributions than unimodal distributions ($P < 10^{-10}$ using likelihood ratio test in all species). A recent paper suggested a 'trimodal' distribution rather than bimodal (WEBER *et al.* 2007). However, the likelihood of bimodal model was also significantly greater than that of trimodal model in all the species analyzed ($P < 10^{-10}$ using likelihood ratio test in all species). Therefore, bimodality of hypo- and hyper- methylated promoters is a common feature in several distantly related vertebrate taxa separated by hundreds of millions of years.

Table 4.1. Normalized CpG contents in promoter and intronic regions of the four vertebrate genomes. Bimodal distribution of low- and high- CpG promoters (LCG and HCG, respectively) is strongly supported based upon statistical tests while introns show unimodal distributions. The mean and medians of CpG O/E are shown.

Species	Number of genes analyzed	Promoter		Intron
		LCG mean (median)	HCG mean (median)	Mean (median) CpG O/E
Zebrafish	4974	0.34 (0.39)	0.83 (0.88)	0.38 (0.38)
Frog	3638	0.21 (0.21)	0.62 (0.63)	0.27 (0.26)
Chicken	2375	0.24 (0.25)	0.83 (0.87)	0.21 (0.22)
Human	7869	0.22 (0.20)	0.74 (0.76)	0.19 (0.17)

It is known that the G and C nucleotide content (G+C content) and CpG O/E are correlated (DURET and GALTIER 2000; FRYXELL and MOON 2005; GARDINER-GARDEN and FROMMER 1987). However, the bimodality of CpG O/E is not caused by GC content bimodality; for example, (SAXONOV *et al.* 2006) showed that G+C contents in human promoter regions follow a Gaussian distribution with one mean. Similarly, we demonstrate that the bimodality of CpG content in promoters is not caused by the underlying distribution of G+C contents (Appendix C and Figure C.3).

To test whether the observed pattern is caused by inaccurate TSS annotation, we restricted our analyses to experimentally verified TSS only, using the data from DBTSS (WAKAGURI *et al.* 2008). There are 4277 human genes in the DBTSS that overlap with our data set. Analyses of this subset of genes show the same results as those from the whole data set (Supplementary Figure C.4). Therefore, our finding is not caused by bias in TSS annotation.

LCGs are associated with tissue-specific genes and HCGs are associated with broadly-expressed genes in distantly related vertebrate genomes.

Next, we investigated the functional implication of the observed bimodality. We first analyzed microarray data from humans to compare with previous results. We analyzed the gene expression data from 79 tissues in gene atlas (SU *et al.* 2004). Genes with LCG promoters were expressed in fewer tissues (median: 38 tissues) than those with HCG promoters (median: 58 tissues). This difference is statistically significant (Table 4.2, Mann-Whitney test, $P < 10^{-3}$) and confirms earlier results (SAXONOV *et al.* 2006). Analysis of exon array expression data (KAPUR *et al.* 2007; XING *et al.* 2007) from six tissues yielded similar results (Appendix C and Figure S5).

To determine whether the same pattern holds in other species, we chose to analyze EST data because they are available from all the species studied here, allowing a meaningful comparison. We obtained EST data from the Unigene database (WHEELER *et al.* 2008). We first compared 100 genes with the highest promoter CpG O/E (Top 100; Table 4.2) and the lowest promoter CpG O/E (Bottom 100; Table 4.2). The top 100 had significantly broader expression than the bottom 100 (Table 4.2, Mann-Whitney test, $P < 10^{-6}$ in all species). Second, we compared the median expression breadths of genes within LCG and HCG classes (obtained by dividing the data where the two Gaussian curves intersected). Genes associated with LCG promoters were expressed in significantly fewer tissues than those with HCG promoters (Table 4.2, Mann-Whitney test, $P < 10^{-3}$ in all species). Furthermore, the relative frequency of HCG increases with the expression breadth in all the four species analyzed (Figure 4.2).

(VINOGRADOV 2005) has shown that intronic CpG content influences expression breadth, potentially as much as the absence and presence of promoter CpG islands (which is equivalent to the distinction between LCGs and HCGs). To gauge whether promoter bimodality affect expression breadth independent of intronic CpG content, we divided human genes into low and high intronic CpG groups (based on median intronic CpG O/E of 0.17). In each group, LCG genes are expressed in significantly fewer tissues than HCG genes (Table C.2). Notably, within each promoter class (LCG or HCG) the expression breadth of genes with low and high intronic CpG content are almost identical (Table C.2). Thus, HCG genes tend to be expressed in more tissues on average than LCG genes, and this distinction is independent of intronic CpG content.

Table 4.2. Expression breadths of genes with low and high CpG content in upstream regions. Results from zebrafish, frog, chicken are from EST database. For humans, results from EST and microarray are both presented.

Species	Number of tissues analyzed	Median number of tissues expressed (proportion of expressed tissues to all tissues)			
		Bottom 100	Top 100	LCG	HCG
Zebrafish	14	3 (21.4%)	6 (42.8%)	4 (28.5%)	5 (35.7%)
Frog	21	4 (19.0%)	8 (38.0%)	4 (19.0%)	6 (28.5%)
Chicken	18	5 (27.7%)	9 (50.0%)	5 (27.7%)	8 (44.4%)
Human	49	7 (14.2%)	33 (67.3%)	10 (20.4%)	30 (61.2%)
Human (microarray)	79	31 (39.2%)	70 (88.6%)	38 (48.1%)	58 (73.4%)

DISCUSSION

We have shown that the structural and functional distinctions between LCGs and HCGs are conserved in several distantly related vertebrate genomes. This indicates that the functional role of DNA methylation in gene expression is conserved in these taxa, which were separated for hundreds of millions of years (Figure 4.1). What is the underlying mechanistic basis of the association of LCG promoters with tissue-specific genes and HCG promoters with broadly-expressed genes in these vertebrate genomes? As mentioned earlier, DNA methylation can suppress gene-expression directly by interfering with transcription factors or indirectly by recruiting chromatin modification enzymes. A recent study (VINOGRADOV 2005) showed that in the human genome CpG content is negatively correlated with chromatin condensation potential, suggesting that LCG promoters will be highly condensed in germline, rendering the associated gene unsuitable for transcription. Thus, vertebrates may have adapted promoter DNA methylation to epigenetically suppress somatic-tissue specific genes in germline. Indeed, we found that in the human genome, promoter CpG content is strongly positively correlated with germline (testis) expression level (Figure 4.3; Spearman's correlation coefficient = 0.38; $P < 10^{-6}$). The median germline expression level of LCG genes is ~ 6-fold lesser than that of HCG genes (Figure 4.3).

The invertebrate chordates with mosaic genomic DNA methylation pattern analyzed here exhibit a single class of hypo-methylated promoters (Figure 4.1, Supplementary Figure C.2). This finding, along with the observation that the structural

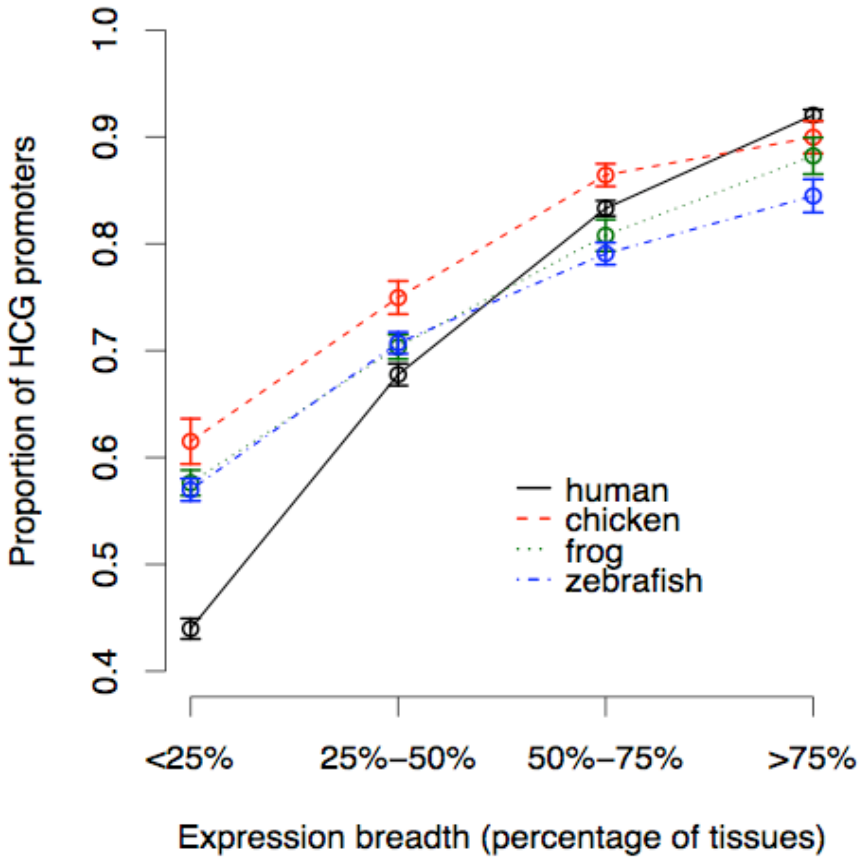


Figure 4.2. Relative frequency of HCG promoters increases with expression breadth in vertebrate genomes. For each vertebrate species analyzed, the genes were divided into four bins based on the percentage of tissues in which they are expressed. Within each bin, the proportion of genes with HCG promoters is plotted. The total number of tissues analyzed in each species is shown in Table 4.2

and functional bimodality in promoters is conserved across several distantly related vertebrate species spanning zebrafish to human, suggests that the bimodality originated early in vertebrate evolution, potentially as a consequence of the transition from mosaic to global methylation of genomes. Hypo-methylation of promoters, as found in *Ciona,* is likely to be the ancestral state because the mosaic DNA methylation pattern in *Ciona* is typical of methylated chordates (SUZUKI *et al.* 2007), and other invertebrate genomes such as those of arthropods and nematodes are free of DNA methylation (TWEEDIE *et al.* 1997). In other words, LCG promoters are a derived feature of vertebrate genomes.

The fact that CpG contents of LCGs are similar to that of the rest of the genome while HCGs preserve CpG contents in several distantly related vertebrate genomes (Figure 4.1 and Table 4.1; also see Supplementary text in Appendix C) provides a clue to the origin of the vertebrate LCG promoters. Specifically, it indicates that the level of DNA methylation in LCG promoters is similar to the genome-wide level. We propose that LCG promoters, and consequently the bimodal distribution of CpG contents in vertebrate promoters, have originated due to mutational decay of CpG dinucleotides following DNA methylation. Our functional analyses suggest that this process has occurred preferentially in upstream regions of tissue-specific genes.

HCG promoters, on the other hand, maintain high CpG contents despite the global genomic methylation. One possible explanation is that broadly expressed genes have selectively avoided DNA methylation and remained as HCGs, because silencing of such genes due to DNA methylation would have been deleterious. For instance, aberrant promoter methylation in the human genome is highly deleterious, often associated with

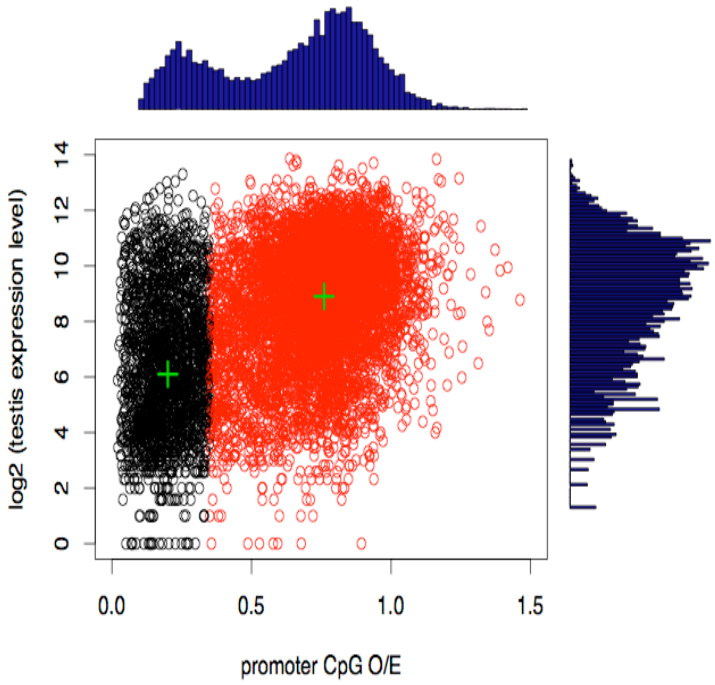


Figure 4.3. Positive relationship between promoter CpG content and germline expression level in human genome. The promoter CpG contents of human genes are plotted against its germline (testis) expression level (in $\log_2$ scale) from exon array data. LCGs and HCGs are colored black and red respectively. The green crosses indicate the median expression level of LCGs and HCGs (6.1 and 8.9, respectively).

several diseases including cancer (EGGER *et al.* 2004; ESTELLER and HERMAN 2002; ROBERTSON and WOLFFE 2000).

According to this model, the rate of CpG loss should be greater in LCG promoters than in HCG promoters. This is indeed the case in the human genome (WEBER *et al.* 2007). Comparing expression patterns in human and mouse genomes has further revealed that CpG islands were preferentially being lost from promoters of tissue-specific genes (JIANG *et al.* 2007). These observations provide strong support to the role of DNA methylation on the origin and evolution of the bimodality of vertebrate promoters.

ACKNOWLEDGEMENTS

We thank comments from the Yi lab, especially from Brendan Hunt. Comments from anonymous reviewers on the previous versions of this manuscript provided helpful insights. This study is supported by funds from the Blanchard-Milliken Fellowship and the Alfred P. Sloan Foundation to S. Yi.

CHAPTER 5

MAMMALIAN GENE REGULATION COMPLEXITY AND THE LENGTH OF CPG ISLANDS

ABSTRACT

Mammalian genomes exhibit a paucity of CpGs due to the dinucleotides' methylation-dependent hypermutability. Certain regions called CpG islands (CGIs), however, escape DNA methylation and are relatively enriched with CpG dinucleotides. CGIs are often associated with protein-coding gene promoters and methylation of CGI is linked with suppression of the associated gene. Several recent studies have divided mammalian promoters into two classes based on the presence or absence of CGI and showed that CGI promoters are associated with widely expressed housekeeping genes, while non-CGI promoters are associated with tissue specific genes. Here we show that the relationship between gene expression and CGIs is more complex. In particular, we show that promoters with long CGIs (LCGI) are associated with genes that exhibit intermediate tissue-specificity, and have on average a larger number of RNA polymerase II binding sites which are occupied in a intermediately tissue-specific manner. Moreover, LCGI promoters are enriched with highly conserved genes involved in important biological processes like development, transcription regulation, and signaling. These results indicate that LCGIs are associated with genes that require more complex gene regulation strategies. The implications of our results on CGI maintenance and regulatory region evolution are discussed.

INTRODUCTION

DNA methylation in mammalian genomes occurs almost exclusively at cytosines immediately followed by a guanine (referred to as "CpG dinucleotides" henceforth). Mammalian genomes are heavily methylated with ~80% of the CpG dinucleotides exhibiting methylated cytosines(BIRD 1980; SUZUKI and BIRD 2008). Methylated cytosines are hypermutable due to spontaneous deamination, which leads to a dearth of CpG dinucleotides in these genomes (FREDERICO *et al.* 1990; FREDERICO *et al.* 1993; FRYXELL and MOON 2005; FRYXELL and ZUCKERKANDL 2000). The human genome, for example, contains only ~22% of the CpG dinucleotides expected from its G+C content (see Chapter 4).

Interestingly, a few regions escape DNA methylation in several tissue types and contain the amount of CpG dinucleotides close to that expected from their G+C content (BIRD *et al.* 1985). These regions, aptly named CpG islands (CGIs), are typically associated with gene promoters and were considered as gene markers in mammalian genomes (ANTEQUERA and BIRD 1993; LARSEN *et al.* 1992). CGIs are known to play crucial roles in gene regulation. Methylation of promoter-associated CGI is linked with transcription suppression, and aberrant CGI methylation is associated with several serious conditions, including cancer (BOYES and BIRD 1991; COMPERE and PALMITER 1981; KASS *et al.* 1997; KESHET *et al.* 2006; KISSELJOVA *et al.* 1998; ROBERTSON and WOLFFE 2000).

Due its importance in mammalian gene regulation, promoter-associated CGIs and their relationship with gene expression have been an area of intense research. Among the findings is the correlation between the presence of CpG islands and gene expression

breadth: studies of human and mouse promoters have shown that promoters with CGIs are associated with broadly expressed genes, whereas promoters with no CGIs are typically tissue specific (CARNINCI 2006; SAXONOV *et al.* 2006). For example, using microarray expression data from 79 tissues, (SAXONOV *et al.* 2006) showed that human promoters with high CpG dinucleotide content, which typically contain CGIs, are associated with housekeeping genes and low CpG content promoters are associated with tissue specific genes. We reported a similar finding in diverse vertebrate genomes, using EST data (Chapter 4). Based on such observations, we have proposed that CpG islands might be selectively maintained to ensure expression of housekeeping genes in germline (Chapter 4).

Here we provide evidence that not only the presence of CGI, but also some characteristics of CGI, in particular the length of CGI plays a role in gene regulation. We show that promoters with longest CGI are not as broadly expressed as genes with typical length CGIs, thus generally exhibit intermediate tissue-specificity. These genes are involved in important biological processes including development, transcription regulation and signaling. According to information theory, genes with intermediate tissue specificity require the most complex mode of gene regulation because of its complex switch-on/off state across tissues. Thus we hypothesized that long CpG island promoter may encode several regulatory signals that are required for a more complex mode of gene expression regulation. We further provide evidence for this hypothesis using RNA polymerase II occupancy pattern.

MATERIALS AND METHODS

Genome sequences and promoter associated CGI annotation

The human genome (version hg 18), mouse genome (version mm9), and CGI annotations were downloaded from the UCSC genome database (KENT *et al.* 2002). The CGI annotation algorithm at UCSC searches the genome sequence one base at a time scoring each dinucleotide (+17 for CpG and -1 for others). Next, it finds maximally scoring segments and annotates it as a CGI if it satisfies the following criteria. 1) G+C content > 50%, 2) length > 200, and 3) the ratio of observed to expected number of CpG dinucleotides (CpG O/E) > 0.6. Because CpG islands thus identified with length < 500 bps are often *Alu* elements instead of real gene control regions (TAKAI and JONES 2002), only the CpG islands > 500bps in length were used in this study. The threshold on base composition (G+C content) and CpG O/E are based on experience. To confirm that the results obtained in our study are not dependent on the CGI annotation algorithm, we re-examined our results using another CGI annotation algorithm that finds CGIs solely based on the CpG dinucleotide property without imposing a base composition a priori assumption (GLASS *et al.* 2007).

For the purpose of finding promoter associated CpG islands, we followed an objective method. Human promoters fall into two classes, the low-CpG content and the high-CpG content promoters (SAXONOV *et al.* 2006). The high-CpG content promoters are often associated with CpG islands. (SAXONOV *et al.* 2006) showed that the average CpG O/E of human promoters peaks at the transcription start site (TSS) and decays gradually as the distance from the TSS increases. When high-CpG content promoters were analyzed separately, we found a similar pattern. Average CpG O/E peaks at the TSS and decays gradually until ~4000 bps on each direction (N.E and S.V.Y unpublished

results). Because this is the average CpG O/E across all high-CpG content promoters, some CpG islands may lay slightly farther that 4000 bps from the TSS. Therefore, we annotated CGIs as promoter associated if they lie within a distance of 5000 bps in each direction around the TSS. Promoters that overlapped with other genes were removed from all the analyses. Alternatively spliced genes were also removed from the analyses because of the uncertainty in TSS position.

Gene expression data

Expression data were obtained from EST counts in the Unigene database (WHEELER *et al.* 2008). Genes with EST count ≥ 1 in a tissue were considered to be expressed in that tissue. The expression breadth of a gene is the number of tissues in which it is expressed. The total number of tissues analyzed in human and mouse are 49 and 47, respectively. We also analyzed exon microarray expression data from six tissues [heart, kidney, liver, muscle, spleen, and testis (XING *et al.* 2007)]. Because of the limited number of tissues from the microarray data, we used tissue specificity index as a measure of expression pattern. Tissue specificity index of a gene is defined as

$$T = \frac{\sum_{j=1}^{n} (1 - [\log_2(Ej) / \log_2(E \max)])}{n - 1}$$

where n is the number of tissues analyzed, Ej is the expression level of the gene in jth tissue, Emax is the maximum expression level of the gene across the n tissues. The higher the tissue specificity index of a gene the more tissue-specific it is.

RNA polymerase II occupancy data

RNA polymerase II (Polr2a) occupancy data were obtained from (BARRERA *et al.* 2008). Briefly, (BARRERA *et al.* 2008) produced a genome-wide map of Polr2a occupancy in five mouse tissue types (brain, heart, kidney, liver and embryonic stem cells) using ChIP-chip. They found ~ 24000 Polr2a binding sites across these tissue types. The relative Polr2a occupancy at each binding site across the five tissues was characterized using Shannon entropy

$$H_s = -\sum_{1 \le t \le N} P_t \log_2 P_t ,$$

where

$$P_t = \frac{B_t}{\sum_{1 \le t \le N} B_t}$$

$\boldsymbol{B_t}$ is the average ChIP-chip log2 ratio in the 1 kbp region centered at the mid-point of the binding site. A high entropy value means Polr2a is bound to that site uniformly across all tissues, whereas a low value of entropy means a more tissue specific binding pattern.

Gene ontology analysis

Gene Ontology analyses were performed using The Database for Annotation, Visualization, and Integrated Discovery [DAVID; (DENNIS *et al.* 2003)]. To find the GO terms overrepresented in genes associated with long CpG islands, all the genes with CpG islands were used as the background and Fishers exact test was performed. The current ontologies in the GO database are molecular function, biological processes, and cellular

component. Analyses were performed on the molecular function and the biological processes domains.

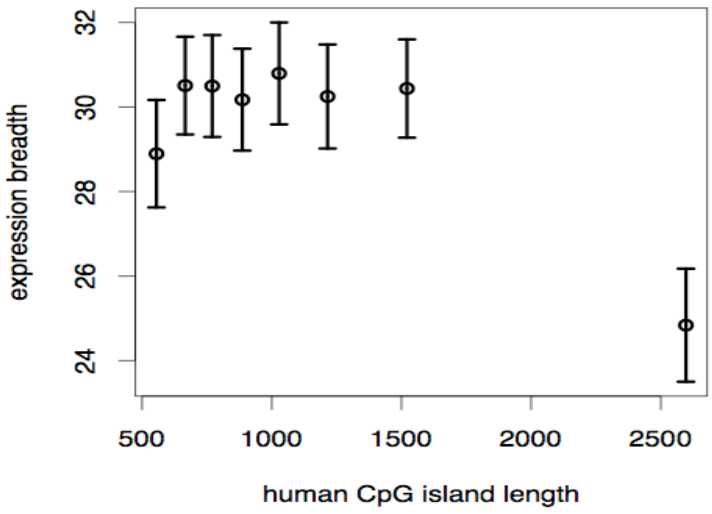
RESULTS

Previous genome-wide studies on the relationship between promoter associated CGIs and the expression pattern of the associated genes often divide the promoters into two classes based on the presence/absence of CGI [Chapter 4, (CARNINCI 2006; SAXONOV *et al.* 2006; VINOGRADOV 2005; WEBER *et al.* 2007)]. These studies have shown that promoters that contain CGIs are associated with housekeeping genes, while promoters without CGIs are associated with tissue specific genes. Interestingly, the length of CGIs varies greatly within genomes (the range of the length of human CGIs in our dataset is 500 -14472 bps). In spite of the importance of CGIs in gene regulation, the reason for such a variation in CGI length and its relationship with gene expression complexity is unknown.

Promoters with long CGIs are associated with genes exhibiting intermediate tissue specificity

We divided the human promoters into equal-sized bins based on the length of the associated CGI and investigated the mean expression breadth of each bin using EST data [Figure 5.1A; (WHEELER *et al.* 2008). This analysis revealed that promoters with long CGIs (LCGIs; CGI length > 2000 bps) are conspicuously different from those with short

A



B

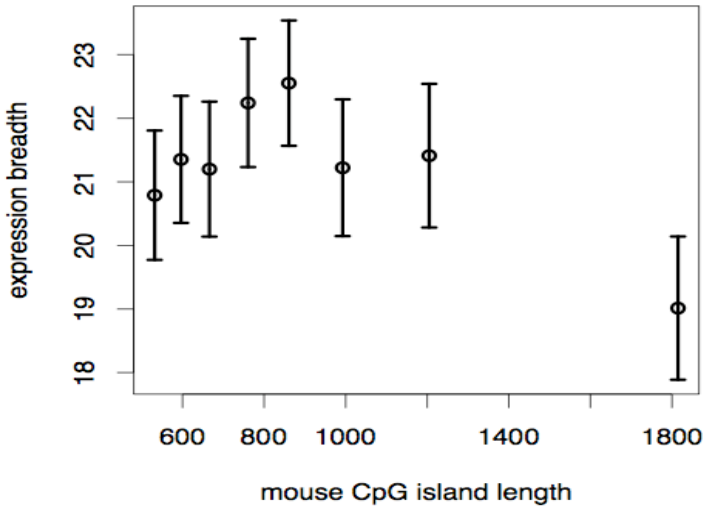


Figure 5.1. Long CpG island promoters are associated with genes expressed in fewer number of tissues compared to genes with short CpG island promoters.
A) Human promoters were divided into eight equal sized bins and the mean expression breadth of the genes in each bin is plotted with its confidence interval. Expression breadths of genes were measured using EST data. B) Same analysis as in A for the mouse genome.

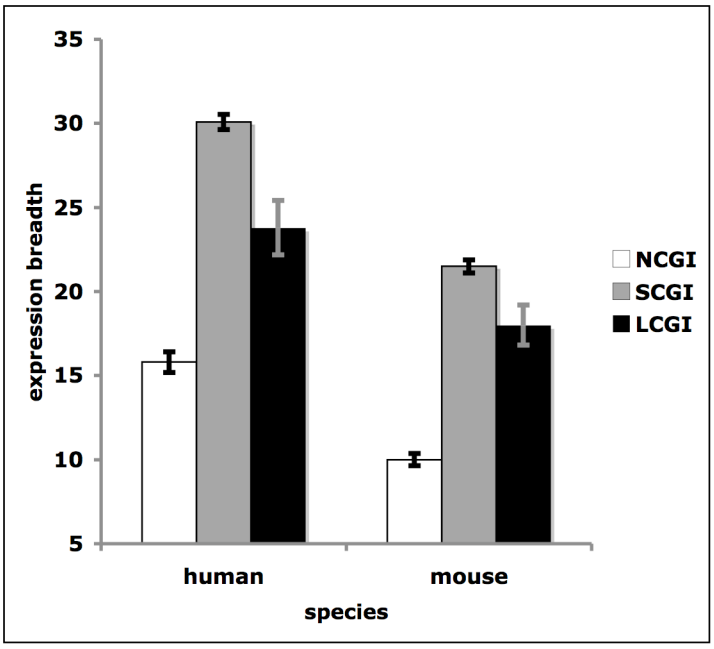
CGIs (SCGIs; CGI length < 2000 bps). Promoters with LCGIs are associated with genes expressed in a fewer number of tissues on average compared to those with SCGIs.

The length of CGIs depends greatly on the annotation algorithm used. The CGI annotations used above are based on an algorithm that uses two criteria to find CGIs. 1) G+C content > 50%, and 2) the ratio of observed to expected number of CpG dinucleotides > 0.6 (See Methods). These arbitrarily chosen thresholds may potentially have an effect on the length of annotated CGIs, which in turn may have confounded our results. To confirm that our results are not a consequence of the annotation method used, we performed the same analysis using another annotation method that depends solely on the clustering property of CpG dinucleotides, the threshold for which is chosen objectively. The results did not change when the new algorithm was used to annotate CGIs (results now shown).

Mouse CGIs are known to be shorter on average compared to human CGIs (JIANG *et al.* 2007). In spite of this difference, the relationship between CGI length and gene expression breadth appears to be conserved between these species (Figure 5.1B). Like in the human genome, mouse promoters with LCGIs (CGI length > 1400 bps) are associated with genes expressed in a fewer number of tissues compared to promoters with SCGIs (CGI length < 1400 bps).

For comparison, we extended our expression breadth analyses to promoters with no CGIs (NCGIs). In accord with previous studies we found that genes with no promoter associated CGIs are more tissue specific compared to genes with promoter associated CGIs (Figure 5.2A). In both human and mouse genomes, the expression breadth of LCGI

A



B

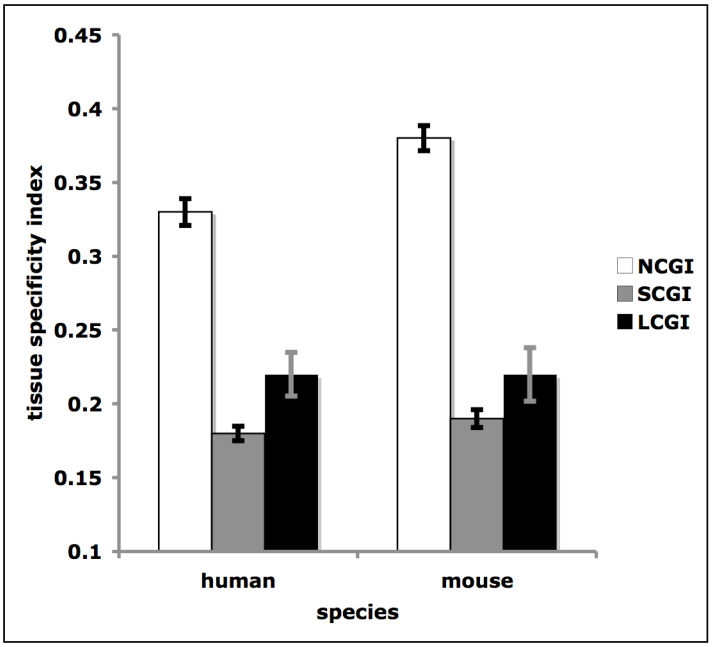


Figure 5.2. LCGI promoters are associated with genes exhibiting intermediate tissue specificity. A) Human and mouse promoters were divided into three groups no CpG island (NCGI), short CpG island (SCGI), and long CpG island (LCGI). See text for the exact definitions of the groups in these species. The mean expression level of genes within each group measured using EST data is plotted with its confidence interval. **B)** The mean tissue specificity index of NCGI, SCGI, and LCGI genes in the human and mouse genomes.

genes are lower than that of SCGI genes and higher than that of NCGI genes (Figure 5.2A).

We also analyzed expression data from six tissues measured using exon microarrays [Figure 5.2B; (XING *et al.* 2007)]. We calculated the mean the tissue specificity indices of genes associated with NCGI, SCGI and LCGI promoters. NCGI and SCGI genes exhibited the highest and the lowest mean tissue specificity indices confirming their highly tissue-specific and broad expression, respectively. LCGI genes exhibited tissue specificity index that is higher than that of SCGI genes but lower than that of NCGI genes. These results suggest that LCGI promoters are associated with genes that exhibit intermediate tissue specificity, which typically require complex gene regulation strategies.

LCGI promoters are complex in terms of Polr2a occupancy

Next, we analyzed the complexity of mouse promoters using genome-wide maps of RNA polymerase II (Polr2a) binding from five tissue types [brain, heart, kidney, liver and embryonic stem cells; (BARRERA *et al.* 2008)]. The median number Polr2a binding site across all mouse promoters is 1. However, the number of binding sites differed when LCGI, SCGI and NCGI promoters were considered separately. We found that the promoters associated with LCGI contain more Polr2a binding sites on average compared to SCGI promoters [median number of binding sites is 2 for LCGI promoters (n= 282) compared to 1 for SCGI promoters (n= 3251); $P < 0.001$ Mann-Whitney test]. LCGI promoters also had a significantly larger number of binding sites compared to NCGI promoters (median number of binding sites is 2 for LCGI promoters (n= 282) compared to 1 for NCGI promoters (n= 964); $P < 0.001$ Mann-Whitney test).

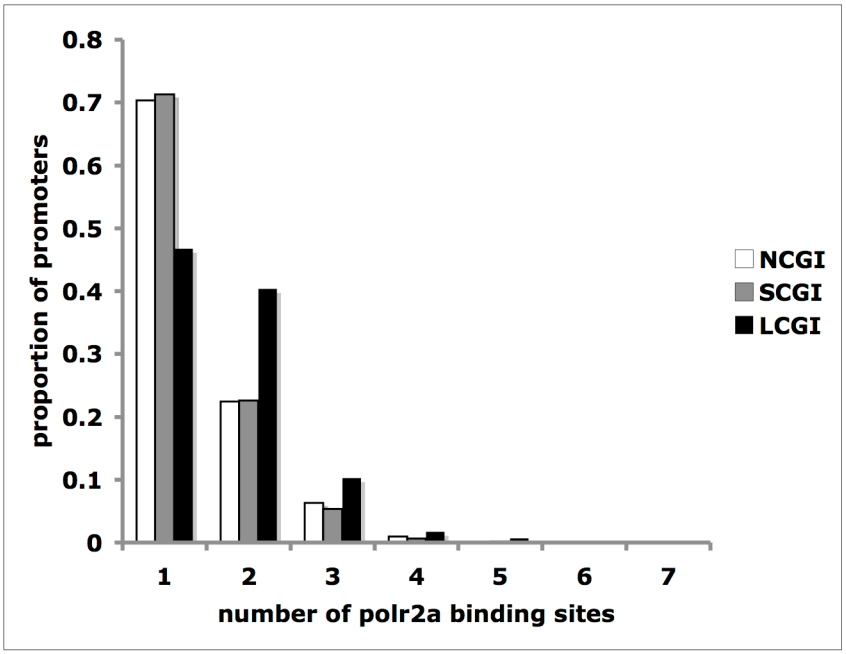
The distinctiveness of LCGI promoters compared to SCGI and NCGI promoters became apparent when the proportion of promoters having a certain number of binding sites was investigated (Figure 5.3A). The distribution in LCGI promoters is clearly different from that in SCGI and NCGI promoters. In both NCGI and SCGI classes ~70% of the promoters contain 1 Polr2a binding site. This measure reduces to ~45% in the LCGI class of promoters. Approximately 21% of the promoters in NCGI and SCGI classes contain 2 Polr2a binding sites, whereas the same measure is ~40% for the LCGI class. LCGI class also contains a greater proportion of promoters with >3 binding sites compared to NCGI or SCGI.

We also measured the average entropy of Polr2a binding sites that overlap with the three classes of promoters (Figure 5.3B). For a binding site, a high entropy value means Polr2a is bound to that site uniformly across all tissues, whereas a low value of entropy means more tissue specific binding pattern. We found that the average entropy of binding sites in NCGI promoters is the lowest, which is in accord with the highly tissue specific expression pattern of NCGI genes (see above). Binding sites in SCGI promoters exhibited the highest entropy, while the entropy of binding sites in LCGI promoters is lower than that of SCGI promoters and larger than that of NCGI promoters. These results indicate that LCGI promoters are more complex compared to SCGI and NCGI promoters in terms of Polr2a occupancy.

LCGI promoters are associated with highly conserved genes involved in important biological functions

The above results indicate that LCGI genes are conspicuously different from SCGI and NCGI genes in terms of gene regulation complexity. It would be interesting to

A



B

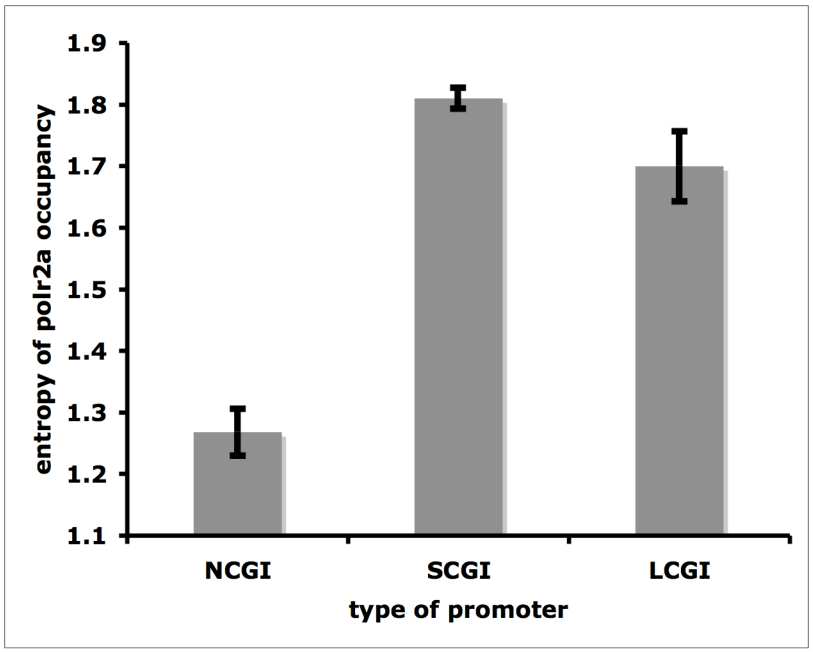


Figure 5.3. LCGI promoters exhibit a more complex Polr2a occupancy pattern compared to SCGI and NCGI promoters. **A)** Mouse promoters were divided into NCGI, SCGI and HCGI classes. Within each class, the proportion of promoters having a certain number (1 through 7) of experimentally verified Polr2a binding sites is plotted. **B)** The average Shannon entropy of Polr2a occupancy in binding sites that overlap with the three classes promoters.

know if the LCGI genes are associated with certain gene functions. We looked for GO terms overrepresented within the human LCGI class using DAVID (DENNIS *et al.* 2003). The top 10 overrepresented GO terms are shown in Table 5.1. In the case of Biological processes, we found that LCGI class is associated mainly with genes involved in development and gene regulation. The top two biological processes GO terms overrepresented in LCGI class are "Development" (Fishers exact test $P < 10^{-14}$) and "Regulation of transcription DNA dependent" (Fishers exact test $P < 10^{-11}$). In terms of molecular function LCGI genes are associated primarily with gene regulation, and signaling. The top two molecular function GO terms overrepresented in the LCGI class are "transcription factor activity" (Fishers exact test $P < 10^{-15}$), and "transcription regulator activity" (Fishers exact test $P < 10^{-14}$).

We also found that human genes with LCGI promoters are more conserved compared to those with SCGI promoters. The median human-chimpanzee dN/dS value of genes with LCGI promoter is 0.046 compared to 0.119 for genes with SCGI promoter (Mann-Whitney test, $P < 10^{-3}$). NCGI genes exhibited the highest dN/dS median value (0.26).

DISCUSSION

Promoter associated CGIs are a common feature of distantly related vertebrate genomes including several taxa like mammals, birds, amphibians, and reptiles (Chapter 4). Several previous studies have shown that CGI promoters are generally associated with housekeeping genes while non-CGI promoters are associated with tissue specific genes in these genomes (CARNINCI 2006; SAXONOV *et al.* 2006; VINOGRADOV 2005; WEBER *et al.* 2007). Because methylation of promoters is linked to suppression of gene expression, it is

Table 5.1: GO terms enriched in genes with LCGI promoters

Biological Processes		Molecular Function	
Overrepresented GO Term	**Fisher's exact test p-value**	**Overrepresented GO Term**	**Fisher's exact test p-value**
Development	1.0×10^{-15}	Transcription factor activity	1.8×10^{-16}
Regulation of transcription, DNA dependent	7.1×10^{-12}	Transcription regulation activity	2.8×10^{-15}
Regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism	1.1×10^{-11}	Sequence-specific DNA binding	9.3×10^{-11}
Regulation of transcription	1.4×10^{-11}	DNA-binding	1.7×10^{-10}
Transcription, DNA dependent	4.4×10^{-11}	Transmembrane receptor activity	1.0×10^{-5}
Regulation of cellular metabolism	1.5×10^{-10}	Signal transducer activity	2.3×10^{-5}
Transcription	1.6×10^{-10}	G-protein coupled receptor activity	4.1×10^{-5}
Regulation of cellular physiological processes	2.1×10^{-10}	Nucleic acid binding	5.3×10^{-5}
Regulation of biological processes	2.4×10^{-10}	Receptor activity	9.8×10^{-5}
Regulation of metabolism	4.9×10^{-10}	Binding	1.9×10^{-4}

proposed that CGI promoters are maintained to ensure expression of housekeeping genes in germline while non-CGI promoters evolved to suppress somatic-tissue specific genes in germline [Chapter 4, (VINOGRADOV 2005)].In these studies, promoters are typically classified into two groups with respect to its association with CpG islands. Here, we show that this dichotomous view of promoters is an oversimplification. Specifically, long CGI promoters are associated with genes that are expressed in a fewer number of tissues compared to short CGI promoters but larger number of tissues compared to no-CGI promoters (Figure 5.2).

What is the reason for the association of LCGI promoters with genes exhibiting intermediate tissue-specificity? Although the general mechanism that maintains CGIs is unknown, several lines of evidence suggest that CGIs may be bound to transcription factors that make them less accessible the DNA methylation machinery (ANTEQUERA 2003). For example, it has been shown that mutations in Sp1 transcription factor binding sites required for mouse *Aprt* gene expression leads to *de nevo* methylation of its CpG island (BRANDEIS *et al.* 1994; MACLEOD *et al.* 1994). Moreover, 5' CGIs that are expected to contain regulatory elements, are typically more resistant to DNA methylation compared to 3' CGIs (BENDER *et al.* 1999).

According to this view LCGIs must be associated with genes that require more complex promoters. Genes that are expressed in intermediate number of tissues may be associated with LCGI promoters because they require complex promoters due to their complex switch-on/off regulation. Our results support this hypothesis by demonstrating that LCGI promoters have a larger number of Polr2a binding sites across five tissues compared to SCGI and NCGI promoters (Figure 5.3A), and the binding sites that overlap

with LCGI promoters exhibit intermediate tissue specificity (Figure 5.3B). Moreover, LCGI promoters are enriched in genes that are involved in important biological processes like transcription regulation and development (Table 5.1). This indicates that these important genes associated with LCGI promoters may use the larger number of binding sites differentially between tissues to achieve the complex switch-on/off pattern required for intermediate tissue specificity.

In light of this link between promoter associated CGI length and gene regulation, it is interesting to note that the length of promoter associated CGIs vary greatly across taxa. CGIs in fish are much smaller than those in mammals (HAN and ZHAO 2008). Within mammals the length of promoter associated CGIs in mouse is on average smaller than that in humans (JIANG *et al.* 2007). Although a strong positive correlation exists between the lengths of CGIs in orthologous promoters from these species the position of CGIs with respect to the TSS is known to vary (JIANG *et al.* 2007). Given that the expression profile of orthologous genes appear to be highly conserved between human and mouse (XING *et al.* 2007), these observations highlight the ability of the regulatory regions to change greatly but still achieve the same expression pattern.

ACKNOWLEDGEMENTS

This study was supported by funds from the Blanchard-Milliken Fellowship and the Alfred P. Sloan Foundation to S. Yi.

CHAPTER 6

CONCLUSIONS

In conclusion, my dissertation encompasses four studies that investigated the evolutionary impacts of DNA methylation on vertebrate genome evolution. In terms of causes, I focused on two important aspects of DNA methylation. First, DNA methylation makes CpG dinucleotides hypermutable due to spontaneous deamination. Second, methylation of promoters is linked with transcriptional silencing of the associated genes. In terms of effects, I focused on neutral and function evolution. The results indicate that DNA methylation has profound impacts on both neutral (see Chapter 2 and Chapter 3) and functional (see Chapter 4 and Chapter 5) evolution of vertebrate genomes.

Chapter 2 investigated the impact of DNA methylation on the neutral molecular clock of primate lineages. The molecular clock hypothesis, which states that the rates of substitution remain approximately equal across evolutionary lineages, has been an area of intense controversy among evolutionary biologists. An opposing hypothesis called the generation-time effect hypothesis, assumes that most of the mutations occur during DNA replication, and states that the rate of substitution should be faster in lineages with short generation-time compared to that in lineages with long-generation time. The study in Chapter 2 uses non-coding sequence data from Old world monkeys and hominoids to show that primate genomes exhibit both time-dependent and generation-time dependent molecular clocks. In particular, methylation dependent substitutions at CpG dinucleotides follow a time-dependent molecular clock, while substitutions at other sites follow a generation-time dependent molecular clock. A mathematical model designed to simulate

the effect of CpG substations on the molecular clock indicated that one potential factor that caused the discrepancy in the results obtained is the proportion of CpG dinculeotides in the dataset.

The results from Chapter 3 indicate that DNA methylation may play an important role in causing intra-genomic variation in substitution rates. A negative correlation was observed between the rate of substitution at CpG dinucleotides and the local G+C content. The strength of this correlation reduces with distance from the CpG dinucleotide and extents up to ~1500-2000 bps. This phenomenon was not observed for non-CpG substitutions, and therefore was proposed to be caused by the local strand separation required for methylation dependent CpG mutations to occur.

Chapter 4 and Chapter 5 investigated the impact of DNA methylation in the functional evolution of vertebrate genomes. Specifically, they focused on the impact of DNA methylation in the evolution of protein-coding gene promoters in vertebrate genomes. Several recent studies have demonstrated that human promoters may be broadly classified into two classes based on the CpG content – the high CpG (HCG) class and the low-CpG (LCG) class. HCG promoters is often contain CpG islands (CGIs). These studies have typically adopted a dichotomous view of promoters (HCG and LCG or CGI+ and CGI-), and demonstrated that the HCG (or CGI+) class is associated with more broadly expressed genes compared to LCG (or CGI-) promoters that are typically associated with tissue-specific genes.

Chapter 4 analyzed several chordate genomes and found that the two classes of promoters is a common feature of several distantly related vertebrate genomes but is not found in an invertebrate outgroup. Moreover, the HCG class of promoters was associated

with more broadly expressed genes compared to those with LCG class of promoters. These observations led to the conclusion that the two classes of promoters potentially evolved early in the vertebrate lineage perhaps as a consequence of the advent of global DNA methylation pattern in vertebrates.

Chapter 5 investigated the promoters with CpG islands in more detail and found that long-CGI promoters are associated with genes exhibiting intermediately tissue-specific expression. Specifically, the tissue-specificity of genes with long-CGI promoters lies between those with short-CGI promoters (widely expressed) and those with no-CGI promoters (highly tissue-specific). An investigation of the complexity of RNA polymerase II occupancy in these types of promoters resulted in the observation that long-CGI promoters contain a larger number of RNA polymerase II binding sites, which are used in an intermediately tissue specific manner. . Moreover, LCGI promoters are enriched with highly conserved genes involved in important biological processes like development, transcription regulation, and signaling. Because genes expressed in intermediate number of tissues may require complex regulatory mechanisms due to their complex switch-on/off state across tissues, these results suggest that long-CGIs are maintained to protect the more abundant regulatory signals in these promoters.

In the post-genomic era with the accumulation of tremendous amounts of epigenetic signal data, we are in great times to explore the consequences of these signals on genomes. My studies have demonstrated that DNA methylation profoundly impacts the structure and function of vertebrate genomes. These studies have enlightened me, and hopefully the scientific community, with the knowledge that this epigenetic signal has played an important role in sculpting the beauty of life on earth at the genomic level.

APPENDIX A

SUPPLEMENTARY INFORMATION FOR CHAPTER 2

Table A.1. Accession numbers of orthologous baboon, chimpanzee and macaque BACs and their locations on the human genome (hg 17; NCBI build 35).

Baboon BAC	Chimp BAC	Macaque BAC	Human Chromosome	Human Start	Human End
AC149010.2	AC158371.2	AC170421.7	chr14	52971380	53198115
AC149109.2	AC155320.2	AC172115.2	chr16	60842337	61130970
AC149627.2	AC166329.3	AC169697.1	chr18	23722496	23969444
AC183349.1	AC166329.3	AC169700.1	chr18	23795853	24091531
AC149841.2	AC157499.2	AC169801.2	chr18	59700968	59948943
AC157860.2	AC175823.2	BX842590.1	chr19	59715046	59920366
AC165404.2	AC165373.2	AC164920.2	chr2	26740092	27065987
AC166617.2	AC169462.2	AC169132.2	chr20	3486153	3745600
AC153752.2	AC153303.3	AC153301.2	chr20	43011761	43235965
AC155785.2	AC159464.1	AC169813.2	chr4	118649444	118926297
AC150015.2	AC159133.2	AC170761.3	chr5	55785355	55990098
AC099743.3	AC161287.2	AC171072.2	chr7	89638064	89864957
AC157859.2	AC157483.2	AC169703.1	chr7	113874528	114116788
AC149461.2	AC159039.2	AC169819.2	chr7	114268826	114509310
AC147705.3*	AC146115.2	AC151364.3	chr7	115347024	115577167
AC149015.2	AC158591.2	AC172195.2	chr8	119184014	119416122

* This region has a partial overlap with ENm001. The portion that overlapped with ENm001 was removed.

Table A.2. Accession numbers for genes used in primate fourfold degenerate site comparison.

Gene	Homo	Macaca	NWM	NWM Genus
ABO	AY268591	AF052080	AY091958	Saguinus
ADCYAP1	S83513	AY775945	AY742807	Saimiri
Aipl1	NM_014336	AF296411	AF296415	Saimiri
ASPM	NM_018136	AY497013	AY497015	Saguinus
CAMP	NM_004345	NM_001033509	DQ471358	Callithrix
CCR5	NM_000579	AF005660	AF452615	Saimiri
CD4	BT019811	D63348	AF452617	Saimiri
CD46	AY916779	U87921	AF025483	Saimiri
CIAS	NM_004895	AY338198	AY338203	Saimiri
CSPG3	AF026547	AY650346	AY665242	Saimiri
DARC	NM_002036	AF311921	AF311918	Saimiri
FUT1	NM_000148	AF080607	AY219617	Saimiri
FUT2	NM_000511	AF136644	AF136645	Callithrix
Hath1	BC069594	AY650376	AY665222	Saimiri
IL10	NM_000572	AB000514	AF294758	Saimiri
IL16	NM_172217	AF017107	AF017109	Saimiri
IRBP	M22453	AJ313476	AF271424	Saimiri
manic fringe	U94352	AY650338	AY665226	Saimiri
MC1R	NM_002386	AY205099	AY205127	Saimiri
MCPH1	BC030702	AY742816	AY570949	Saimiri
MSX1	NM_002448	DQ067471	DQ067473	Saimiri
NGFB	NM_002506	AY091928	AY665218	Saimiri
NTRK3	BC013693	AY742818	AY665246	Saimiri
ODC1	AY841870	AY091984	AY091985	Saguinus
OXTR	NM_000916	U82440	AY091952	Saguinus
PHFAH1B1	AF400434	AB083321	AY665240	Saimiri
POMC	NM_000939	M19658	AY091995	Saguinus
Prp/PRNP	AF085477	AB125193	U08310	Saimiri
PRSS12	NM_003619	AY862980	AY862982	Saguinus
RH30	NM_016124	S70343	AF012429	Callithrix
RNASE1	NM_002933	AF449632	AF449637	Saimiri
SOD2	AY267901	AB201468	AB087281	Callithrix
Sp17	Z48570	AF005551	AF134585	Callithrix
T2R4	NM_016944	AY725015	AB199025	Callithrix
T2R8	NM_023918	AY724999	AB199060	Callithrix
T2R13	NM_023920	AY725009	AB199115	Callithrix
T2R38	AY724960	AY725023	AB199155	Callithrix
T2R40	NM_176882	AY725014	AB199175	Callithrix
TGM4	NM_003241	DQ150493	DQ150498	Saguinus
ZAN	NM_003386	AY428851	AY428855	Saimiri
ZP2	NM_003460	Y10690	Y10767	Callithrix

APPENDIX B

SUPPLEMENTARY INFORMATION FOR CHAPTER 3

SUPPLEMENTARY TEXT

Obtaining the human-chimpanzee-baboon triple alignments for sequences mined from GenBank

We obtained 377 baboon BACs from Genbank (Benson et al. 2006). Blastz (Schwartz et al. 2003) was used to align the baboon BACs to the human genome. The high scoring segment pairs returned by Blastz were converted into PSL format using the lavToPsl program and chained using the axtChain program. Next, the "best chain" for each baboon sequence was obtained using a perl code. To be the best chain of a baboon sequence, a chain must satisfy two conditions: First, it must be the highest scoring chain for that baboon sequence. Second, it must cover at least 60% of the total length of the baboon sequence. We further filtered the best chains such that the score per site is at least 50. There were 274 such chains. These chains were then "netted" against the human genome using the chainPreNet and chainNet programs. The nets were then split based upon the chromosomes and converted into human-baboon alignments (in MAF file format) using the following programs in order netChainSubset (with –splitOnInsert option), chainToAxt, axtToMaf.

A similar procedure was used to obtain human-chimpanzee alignments (the score pre site cutoff for chimpanzee best chains was 80, and there were 922 chains satisfying this condition). The human-chimpanzee and human-baboon alignments were projected on

to the human sequence using the mafProject program, converted to human-chimpanzee-baboon triple alignments using the multiz program. Finally the triple alignments were converted into fasta format using the maf2fasta program.

Higher rates of CpG substitutions in some transposable elements

In our data, transposable elements had a significantly higher rate of CpG substitutions as compared to non-repetitive regions: the rate of CpG substitution is ~14% higher in transposable elements as compared to non-repetitive regions (15.08 ± 0.5%in transposable elements compared to 13.21 ± 0.4%in non-repetitive regions).

However this difference was only significant in intergenic regions. While there was a 21% increase in the rate of CpG substitution of intergenic transposable elements as compared to intergenic-non-repetitive regions (15.64 ± 0.6% in repetitive-intergenic as compared to 12.94 ± 0.6% in non-repetitive intergenic), rates for intronic transposable elements was 14.19 ± 0.8%, not significantly different from those for non-repetitive intronic regions 13.60 ± 0.7%, $P > 0.05$).

This disparity was caused by at least two factors: first, different classes of repetitive elements exhibit different rates of CpG substitutions, and second, that genic (intronic) and non-genic (intergenic) regions harbor different types of transposable elements. Different types of transposable elements exhibited different rates of CpG substitutions. CpG substitution rates for LTRs, LINEs and DNA transposons 18.3 ± 1.9%, 17.5% ± 1.5% and 16.1% ± 3.5% respectively, significantly greater than rates for non-repetitive regions (13.21 ± 0.4%). SINEs however have a lower rate (13.20 ± 0.8%), similar to that of the non-repetitive regions (the difference in CpG substitution rates

among different transposable element classes has previously been observed by Meunier et al. [2005]).

Second, SINEs are more enriched in genic, high GC regions (Lander et al. 2001). In our data, SINEs contribute to 71.8% of all CpG sites in the repetitive portions of introns, compared to 56.1% of CpG sites in intergenic repetitive regions. When SINEs were removed from the data, rates in intronic transposable elements were significantly higher than intronic non-repetitive regions (18.22 ± 1.3% versus 13.60 ± 0.7%).

Previous studies on length effect and CpG substitution rates

As we discuss in the text, other studies including Fryxell and Moon (2005) and Zhao and Jiang (2007) have investigated length effects on rates of CpG substitutions. In particular, Zhao and Jiang (2007) analyzed 292216 CpG and GpC sites from the Human SNP database (http://www.ncbi.nlm.nih.gov/projects/SNP/). They used noninsertion/deletion, non-repetitive, biallelic polymorphic sites for which the sequence data of at least 100 nucleotides flanking each side of the site is known. The flanking sequences were then used to map the polymorphic site to the human genome and extract sequence information for 500 nucleotides on each side of the polymorphic sites. Next, they mapped human sequence to the chimpanzee genome to establish orthology. A parsimony method was then used to infer ancestral alleles and calculate the rate of CpG substitution.

Zhao and Jiang observed that the absolute value of the slope of the relationship between log (CpG substitution rate) and GC content *increased* with the length over which GC content was measured (segment lengths of 101 bps to 1001 bps around the CpG site,

Figure 1C in Zhao and Jiang 2007). This result contradicts our results. Furthermore, Zhao and Jiang (2007) do not provide an explanation for this increase in slope.

However, there are several advantages of our methodology compared with that of Zhao and Jiang (2007). First, we use a more accurate approach for inferring CpG and GpC substitution rates. For example, we removed GpCs that may overlap with CpGs (and vice versa). Moreover, we use a better dataset. They use the SNP database, which in contrast to data obtained by resequencing methods, suffers from ascertainment bias.

Their analyses may have been influenced by the drawback presented by the data or other errors in methodology. For example, Zhao and Jiang (2007)'s CpG and GpC rates are startlingly large. Specifically, in segments of 101 nucleotides, they report CpG rates over 0.7 whereas other studies report rates less than 0.2 (Nachman and Crowell 2000; Fryxell and Moon 2005; Meunier et al. 2005; Kim et al. 2006; Taylor et al. 2006), and they report GpC rates over 0.3, which is over an order of magnitude larger than other studies (Fryxell and Moon 2005).

A second issue with the analysis of Zhao and Jiang (2007) is their use of overlapping windows. In particular, they consider the effect of the G+C content in the flanking 101 nucleotides, and then contrast this with, for example, the effect of the G+C content in the flanking 501 nucleotides. However, the closest 101 nucleotides is being used in both calculations, which confounds its effect. In contrast, we use a sliding window analysis to reduce any confounding effects.

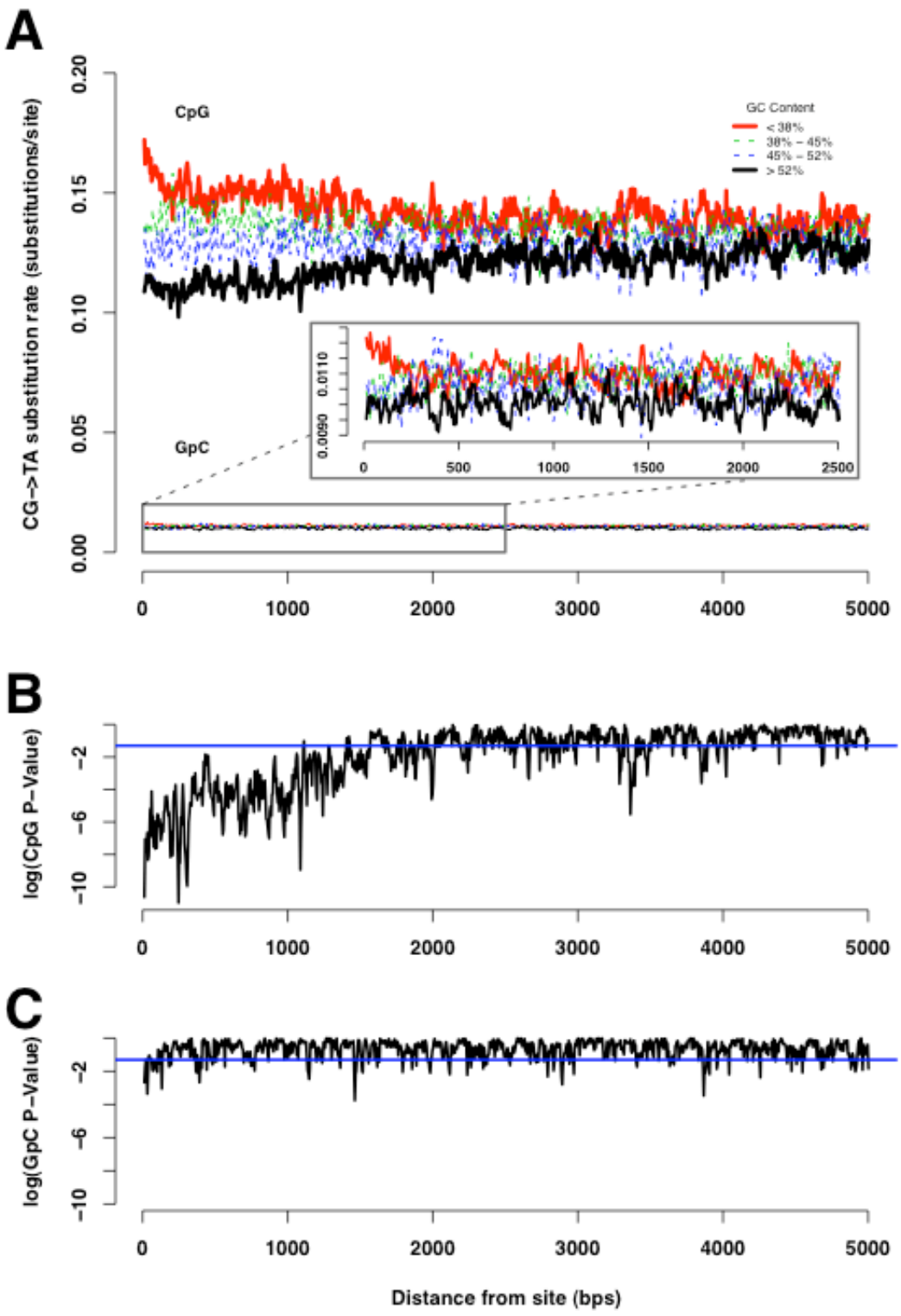


Figure B.1. Sliding window analysis of the relationship between CpG substitution rate and normalized G+C content. The same experiment as in Figure 3 with window

size 25 and step size 5. **(A):** The distance decaying effect of G+C content on the rate of CpG substitution persists even with a smaller window size of 25 bps (as compared window size of 200 bps in Figure 3. 3). In the case of GpC sites, there was no distance decaying effect**. (B):** Results of the chi-square test for the independence of the rate of CpG substitution and the G+C content of the windows. The blue line indicates $\log_{10}$(*P*-value) = -1.30. The distance decaying effect subsided after ~2000 bps **(C):** Results of the same experiment as in panel B, but for GpC sites. There is no distance decaying effect.

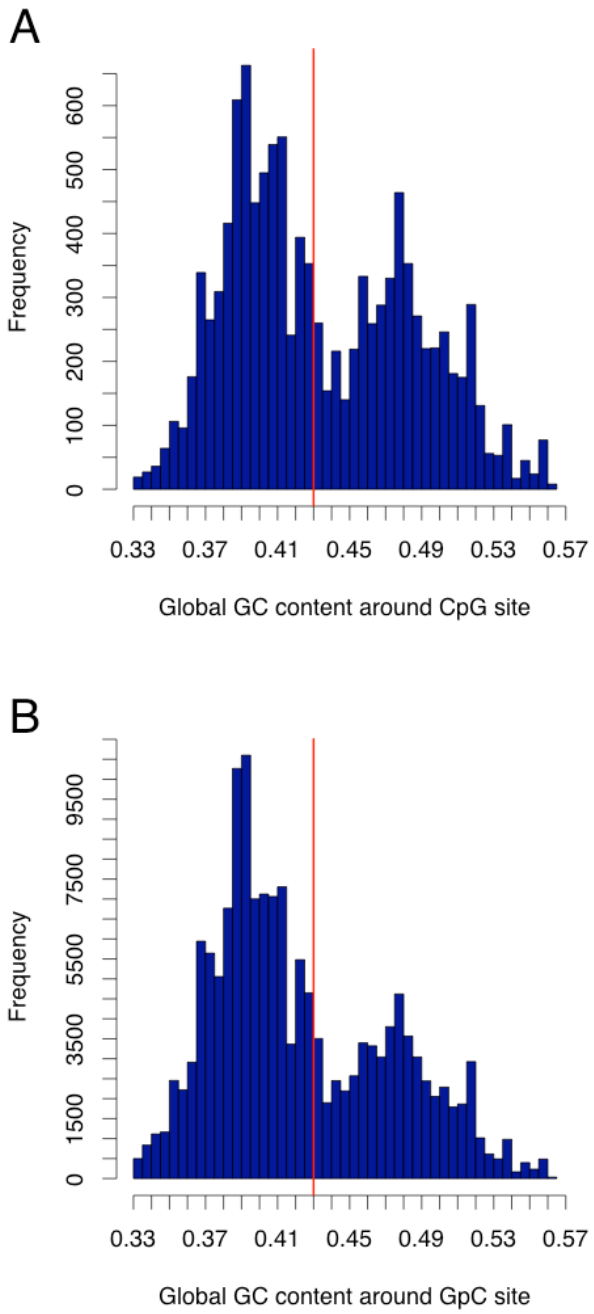


Figure B.2. The distribution of GC content in 100kb segments around CpG and GpC sites. (A) The G+C content of 100kb segments around CpG sites GC_{global} followed a bimodal distribution with means 39% and 48% respectively. The red line indicates GC_{global} = 43%, which was used as the cutoff to differentiate between low- GC_{global} and high-GC_{global} regions. **(B)** G+C content of 100kb segments around GpC sites also exhibited a bimodal distribution, with approximately the same means as those of GC_{global}. The red line marks G+C content of 43%.

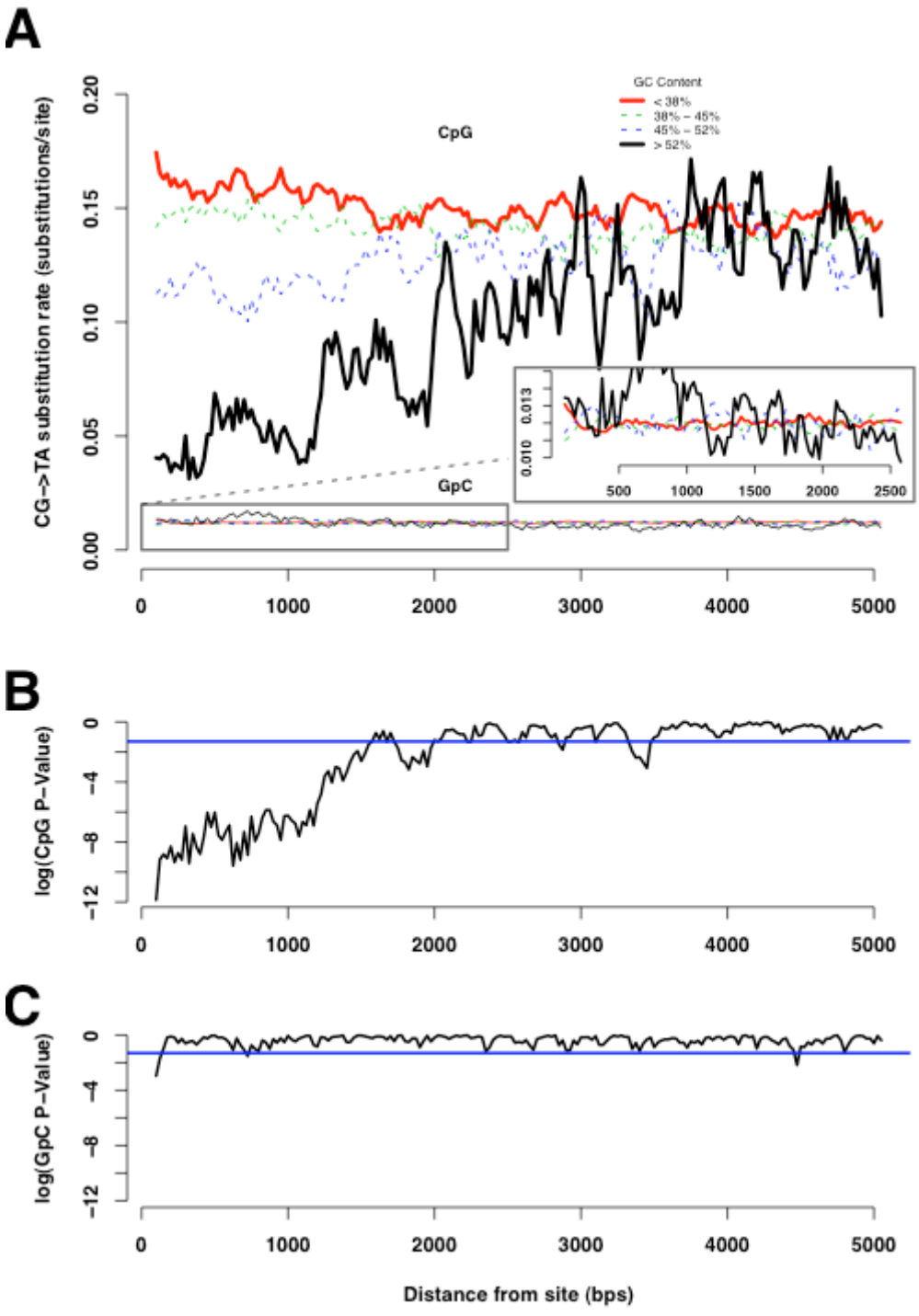


Figure B.3. Distance decaying relationship between G+C content and CpG substitution rate in low- GC_{global} regions. Same analysis as in Figure 3 in the paper with

CpG (and GpC) sites with G+C content of 100kb segments around them less than 43%. **(A)** The distance decaying effect of local G+C content on the rate of CpG substitutions was apparent, and the curves converged at ~1500 bps. In case of GpC, there was no distance decaying effect. **(B)** The test for independence of G+C content and the rate of CpG substitutions. The distance decaying effect was apparent from the gradual increase of *P*-values with increase in distance. P-Values become insignificant at ~ 1500 bps. **(C)** The results of the test for independence of G+C content and the rate of CpG substitutions. No distance decaying effect was observed between GpC substitution rate and G+C content.

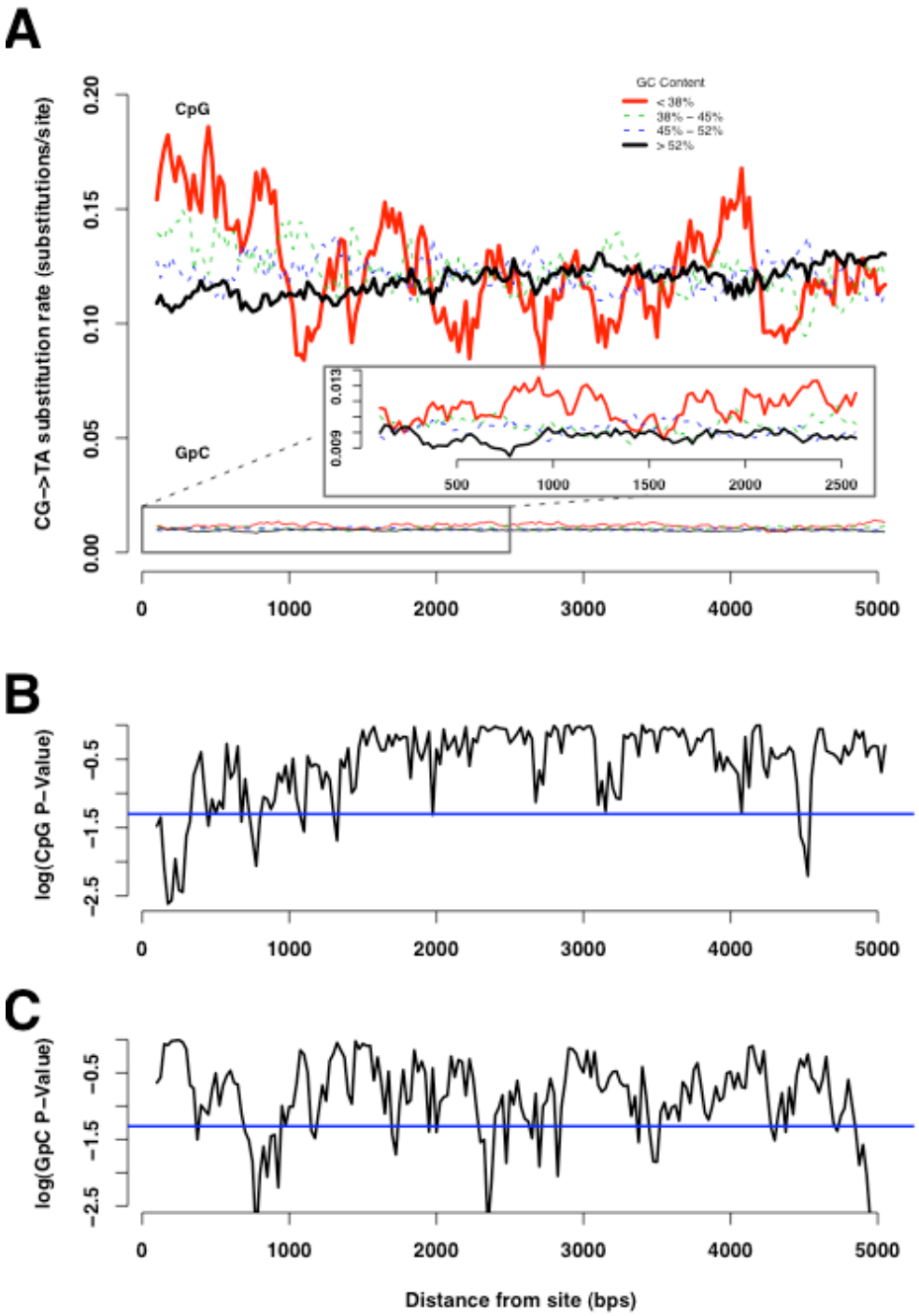


Figure B4. Relationship between G+C content and CpG substitution in high-GC_{global} regions. Same analysis as in Supplementary Figure 2 with CpG (and GpC) sites with G+C content of 100kb segments around them greater than 43%. **(A)** The distance

decaying effect for CpG sites was not apparent because of the fluctuations caused by reduced sample size in bins. In case of GpC, there was no distance decaying effect. **(B)** The test for dependence of CpG substitution rate and G+C content was insignificant starting at distances close to the CpG site. **(C)** No distance decaying effect was observed between GpC substitution rate and G+C content.

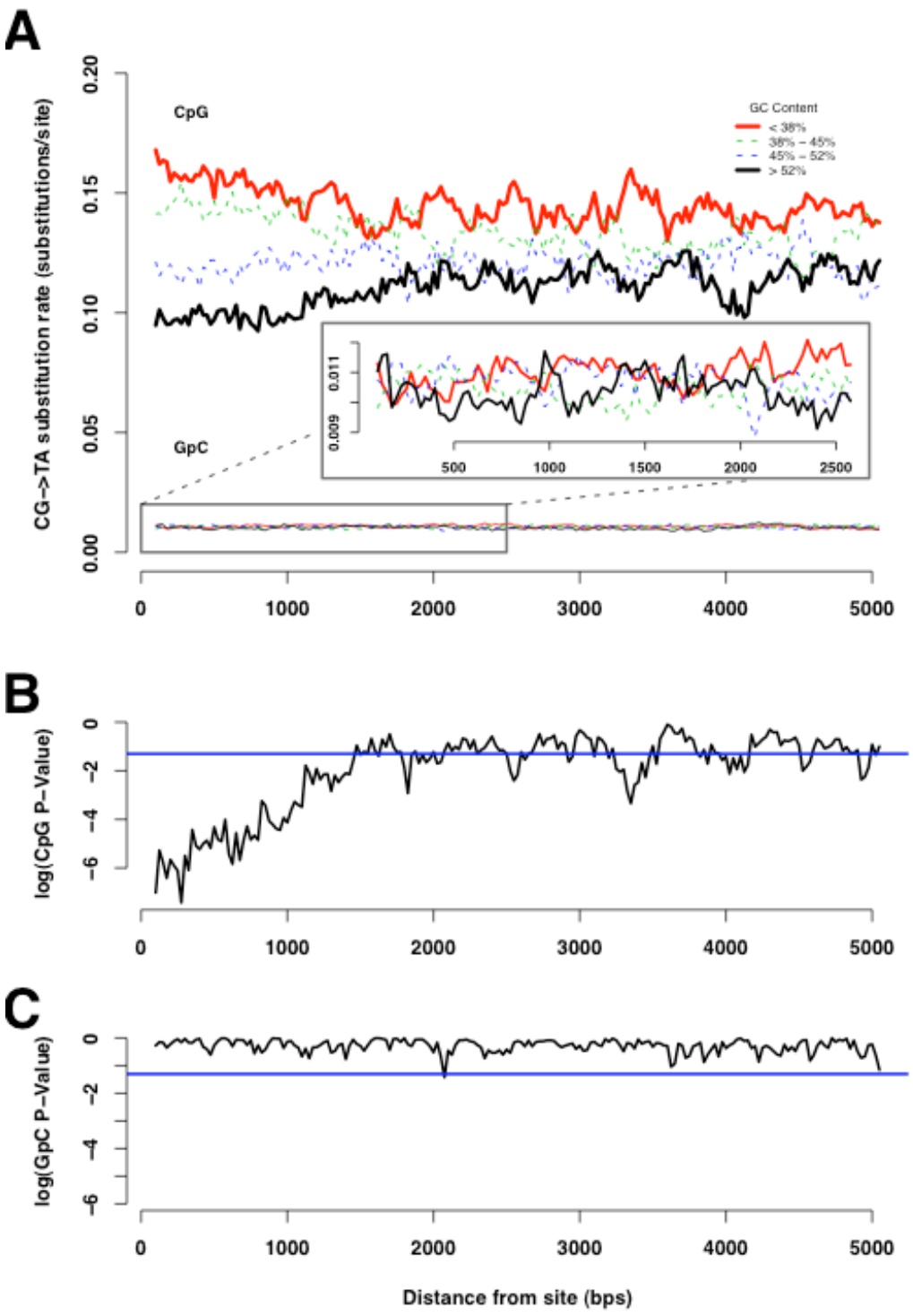


Figure B.5. Relationship between G+C content and substitution rate of randomly picked CpG and GpC sites. Same analysis as in Figure 3. 3 of the paper with CpG and GpC sites randomly picked from low- GC_{global} and high- GC_{global} regions. A total of 6000

(3000 each from low- GC_{global} and high- GC_{global} regions) randomly picked CpG sites and 70000 (35000 each from low- GC_{global} and high- GC_{global} regions) randomly picked GpC sites were used in this analysis. The results obtained were similar to that obtained in Figure 3. 3 and Figure 3. 4 in paper, suggesting that the distance decaying relationship is not a consequence of overrepresentation of CpG sites from low G+C- or high G+C-content regions.

.

APPENDIX C

SUPPLEMENTARY INFORMATION FOR CHAPTER 4

SUPPLEMENTARY TEXT

Analysis of human intergenic regions

We analyzed the distribution of CpG O/E in intergenic regions of the human genome. Intergenic regions were defined as regions that are at least 5000 bps away from any gene in the "UCSC gene" and "Ensembl gene" annotation tracks in the UCSC genome browser (Karolchik et al. 2008). The "UCSC gene" track consists of genes from RefSeq, GenBank and Uniprot. The "Ensembl gene" track consists of genes from the ensembl gene annotation pipeline. There were 15378 intergenic regions with medial length of 29449 bps. The distribution of intergenic CpG O/E (Figure S1) is similar to that of intragenic CpG O/E (Figure 1 in main text). The median intergenic CpG O/E is 0.17, indicating heavy methylation.

Distribution of promoter CpG O/E and intragenic CpG O/E for species with inaccurate gene annotations

For several species RefSeq gene annotations (Pruitt, Tatusova, and Maglott 2007) were not available or very few RefSeq genes were available (e.g., *Strongylocentrotus purpuratus*). The results from these species are presented here. When RefSeq genes are not available, we used gene annotations from Ensembl (Hubbard et al. 2002), except in

Table C.1: Names, genome builds, and gene annotation information of all the species in the current study.

Species	Genome Build	Gene Annotation (# genes analyzed)
Sea Squirt (*C. intestinalis*)	ci2	Refseq (525)
Sea Squirt (*C. Savignyi*)	Csav2	All Ensembl Genes* (8,955)
Sea Urchin (*S. purpuratus*)	strPur2	Refseq (131)
Zebrafish (*D. rerio*)	danRer4	Refseq (4,974)
Fugu (*T. rubripes*)	fr2	Ensembl Known Genes ** (102)
Frog (*X. tropicalis*)	xenTro2	Refseq (3,638)
Lizard (*A. carolinensis*)	anoCar1	Mapped Human Genes (1,450)
Chicken (*G. gallus*)	galGal3	Refseq (2,375)
Human (*H. sapiens*)	hg18	Refseq (7,869)

***** The ensemble annotation has < 35 known genes for this species. Therefore, all the ensembl genes (which include mainly *predicted* genes) were used for the analysis.
** Only single transcript genes with a 5' UTR were used for this species.

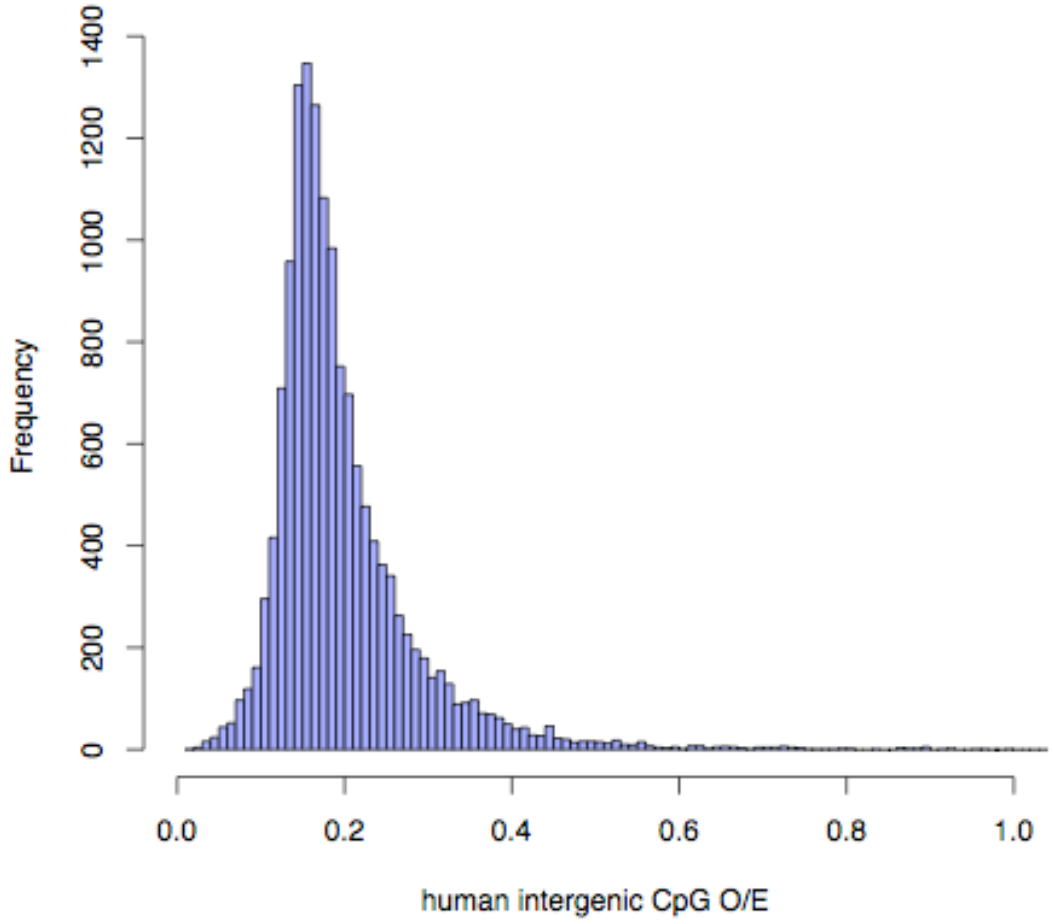
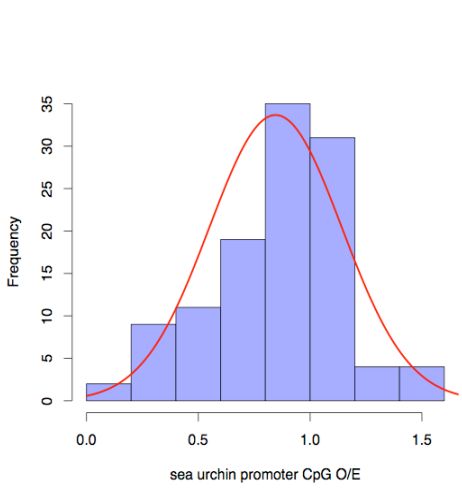
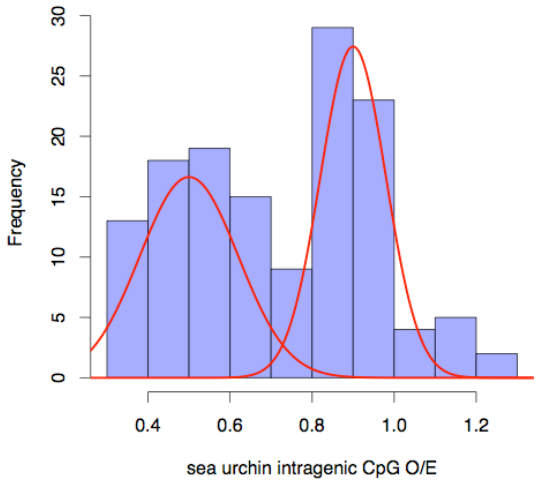


Figure C.1. Distributions of CpG O/E of intergenic regions of the human genome. Intergenic regions are defined as regions at least 5000 bps away from any UCSC or Ensembl genes. The median human intergenic CpG O/E is 0.17.

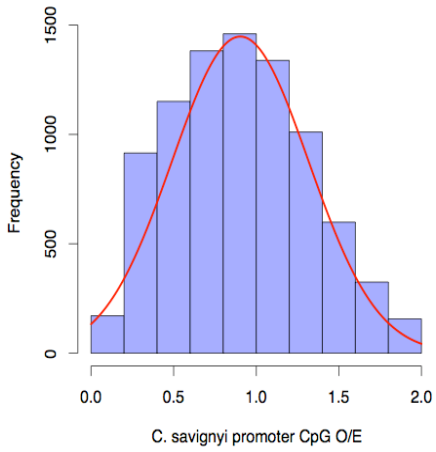
the case of lizard where we used human genes mapped to the lizard genome (Supplementary Table 1). The results from invertebrate genomes *S. purpuratus* and *C. Savignyi* (Supplementary Figure S2) were similar to that from *C. intestinalis* (Figure 1 in main text). Results from lizard (Supplementary Figure S2) were similar to that from other vertebrates (Figure 1 in main text). Although the fugu genome is globally methylated, some previous studies failed to detect CpG clustering in its genome (eg. Glass et al. 2007). In contrast, we found that the bimodal distribution is a better fit to the promoter CpG O/E distribution than the unimodal distribution (Supplementary Figure S2). However, considering the different methods and annotations used by the previous studies and inaccurate annotation used here, these results must be taken with caution. We stress that even the absence of a bimodal distribution of promoter CpG O/E in fugu genome does not counter our proposal that bimodality emerged early in vertebrate evolution. Given the fact that distantly related vertebrate taxa (including zebrafish) exhibit a bimodal distribution of promoter CpG O/E (Figure 1 in main text), it is more likely that fugu has either lost bimodality or the bimodality is undetectable because fugu promoters are much smaller than the 600 bp region considered in this study. Given the compact genome of fugu (Aparicio et al. 2002), the latter possibility appears to be more likely. The fugu genome clearly requires a more detailed analysis, perhaps when more accurate, experimentally verified transcription start site annotations become available.

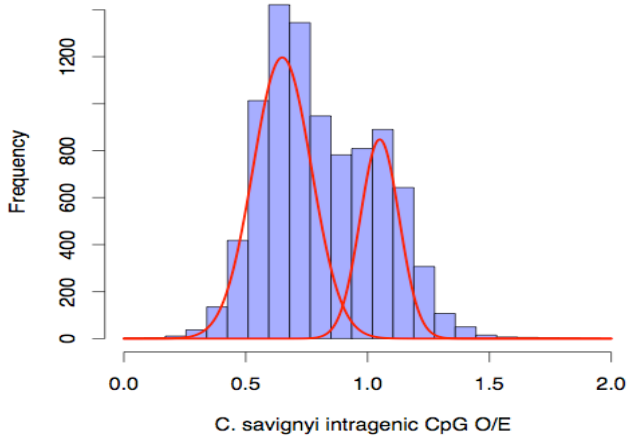
Sea Urchin (*Strongylocentrotus purpuratus*)



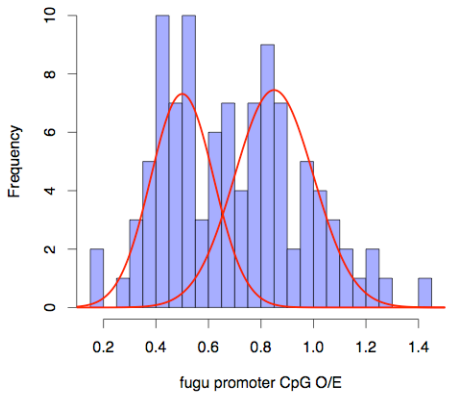


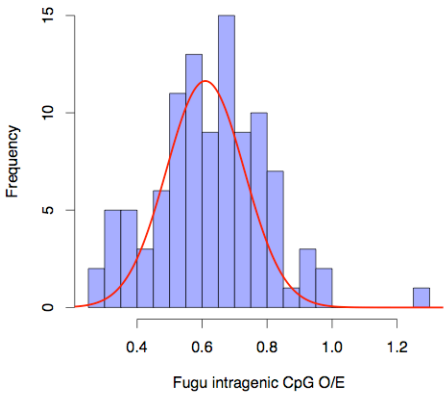
Sea squirt (*Ciona savignyi*)



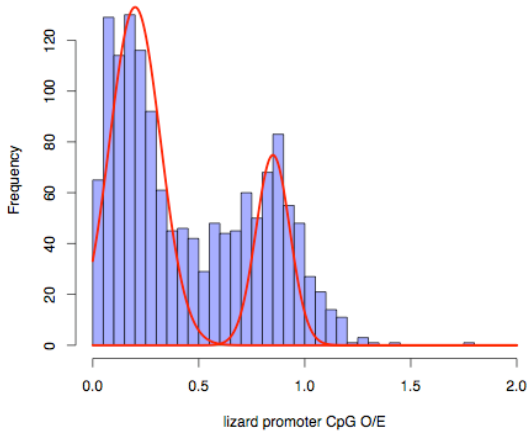


Fugu (*Takifugu rubripes*)





Lizard (*Anolis Carolinensis*)



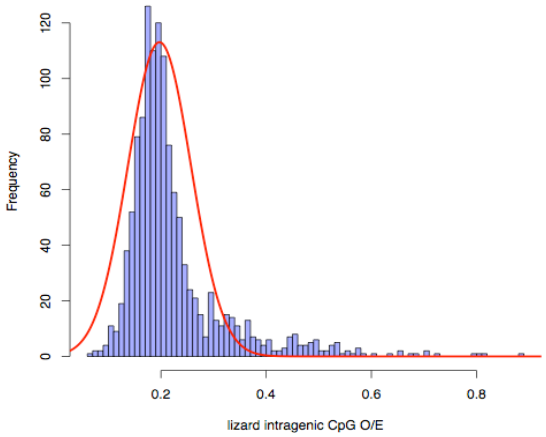


Figure C.2. Distributions of CpG O/E of promoters and intragenic regions of *C. savignyi*, *Strongylocentrotus purpuratus, Takifugu rubripes*, and *Anolis Carolinensis.* The definition of promoters and intragenic regions are the same as in main text. The red curves indicate the best fitting normal curves for each of these distributions.

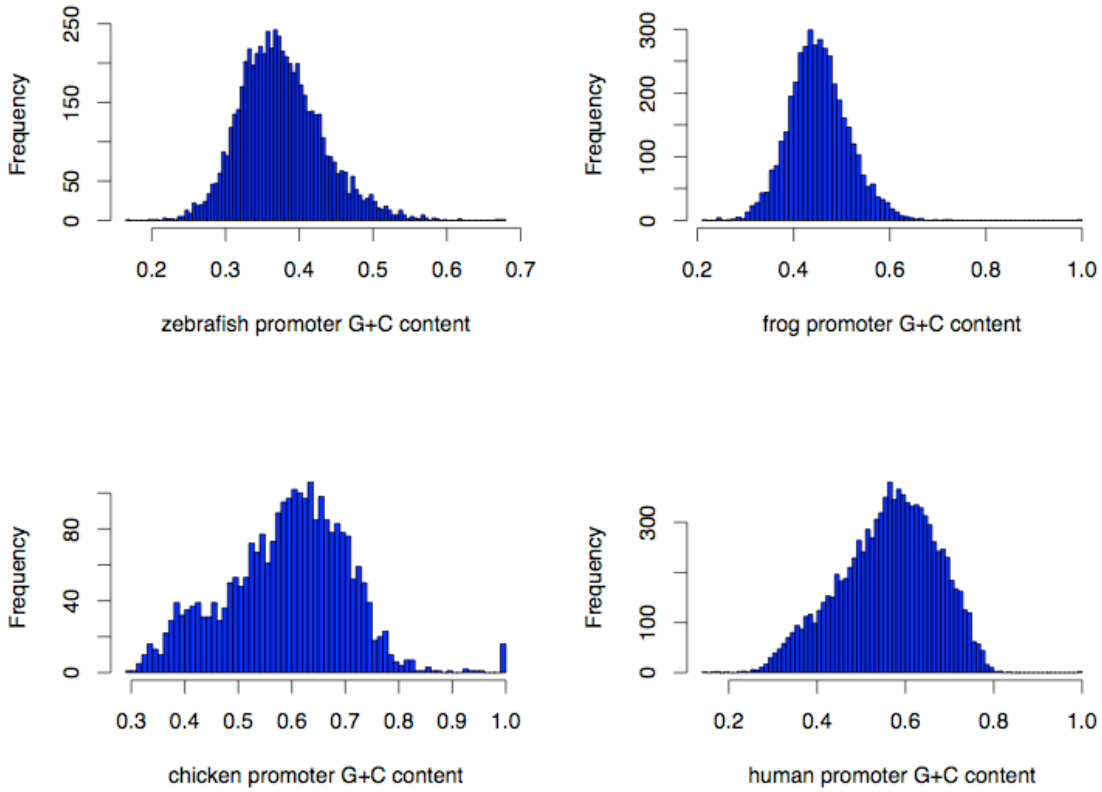
Bimodal distribution of promoter CpG O/E in vertebrates is not caused by underlying G+C Content

To test if the bimodal distribution of promoter CpG O/E is caused by G+C content, we first plotted the histogram of promoter G+C content in zebrafish, frog, chicken and human (Supplementary Figure S3A). The distribution of G+C content in zebrafish, frog, and human are clearly unimodal, indicating that the bimodal distribution of CpG O/E is not caused by G+C content in these species. Only in the case of chicken, the distribution was not clearly unimodal. However, a scatter plot of promoter G+C content vs. CpG O/E (Supplementary Figure S3B) shows that a major fraction of low CpG content promoters (LCG class) exhibit high G+C content (greater than 0.5) in both human and chicken genomes. This indicates that G+C content is not the cause of bimodal CpG O/E distribution in these species.

Comparison of CpG content in introns and LCG promoters

It should be noted that although CpG O/E of LCGs are similar to but not exactly the same as the intronic CpG O/E (Table 1 in main text). This can be caused by two factors: first, we used a small number of nucleotides to estimate CpG O/E from promoters (600 bps), while the numbers of intronic sites used are much larger. The difference in CpG O/E may simply reflect statistical fluctuation. Alternatively, additional selective mechanisms may be at work to fine-tune CpG contents of LCG promoters. For example, to use methylation as a regulatory tool, some CpG sites should be preserved, and this can influence the level

A


Frequency
0
50
150
250
0.2
0.3
0.4
0.5
0.6
0.7
zebrafish promoter G+C content
Frequency
0
100
200
300
0.2
0.4
0.6
0.8
1.0
frog promoter G+C content
Frequency
0
40
80
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
chicken promoter G+C content
Frequency
0
100
300
0.2
0.4
0.6
0.8
1.0
human promoter G+C content

B

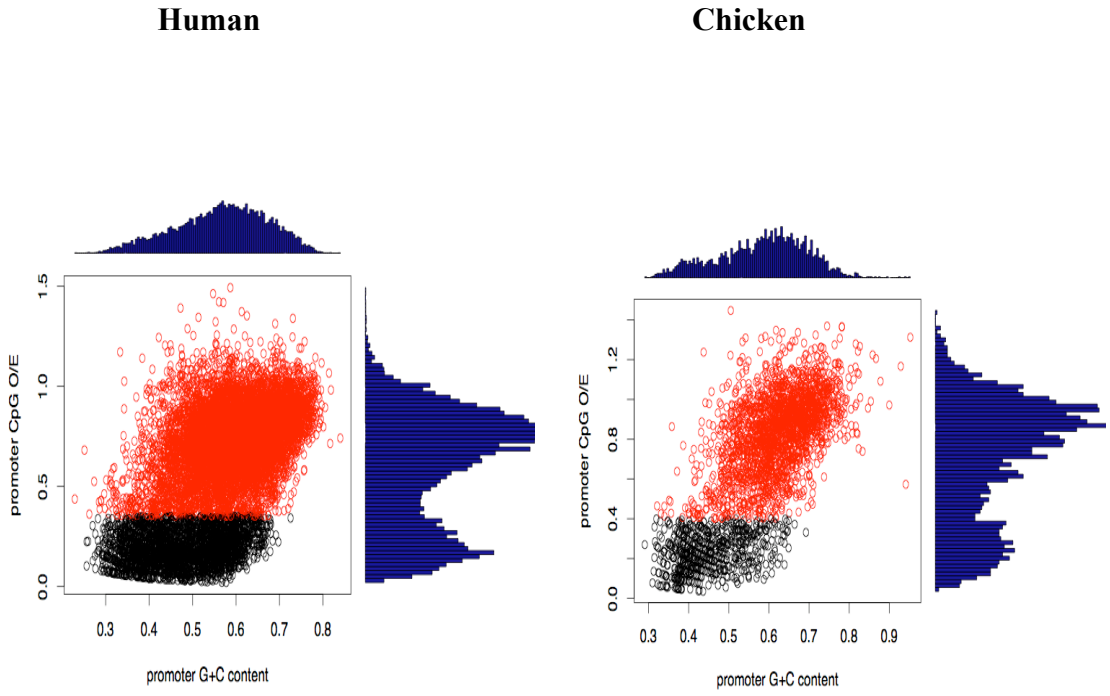


Figure C.3. Bimodality of vertebrate promoters CpG O/E is not due to the G+C contents. A) Distribution of promoter G+C content in zebrafish, frog, chicken and human genomes. **B)** Scatter plot of promoter G+C content and promoter CpG O/E for human and chicken genomes. LCG and HCG promoters are colored black and red, respectively.

of CpG contents. In the current data set however, there was no consistent trend in terms of the relative level of normalized CpG contents in introns and LCGs across the four vertebrate genomes (Table 1 in main text).

Analysis of human exon array expression data

For the human genome, we analyzed exon array expression data from 6 tissues (Xing et al. 2007). Gene expression levels obtained from exon arrays are more accurate (Kapur et al. 2007; Xing et al. 2007) compared to that from 3' arrays data used in gene atlas (Su et al. 2004). Because of the limited number of tissues for which the expression data is available, we used tissue specificity index (Yanai et al. 2005; Liao and Zhang 2006) as a measure of expression pattern. Tissue specificity index of a gene is defined as

$$T = \frac{\sum_{j=1}^{n} (1 - [\log_2(Ej) / \log_2(E \max)])}{n - 1}$$

where n is the number of tissues analyzed, Ej is the expression level of the gene in jth tissue, Emax is the maximum expression level of the gene across the n tissues. The higher the tissue specificity index of a gene the more tissue-specific it is. We found a strong negative correlation between promoter CpG content and tissue specificity index ($r^2 = 0.82$; $P < 10^{-3}$), indicating that LCG genes are much more tissue specific than HCG genes.

Table C.2. Median expression breadth of HCG and LCG genes with low and high intronic CpG O/E. Human genes were divided into two groups based on the median intronic CpG O/E of 0.17. Within each group, the median expression breadth of genes with HCG and LCG promoters are shown.

	Median expression breadth*	
Intron CpG O/E	LCG promoters	HCG promoters
Low (<= 0.17)	10	30
High (> 0.17)	11	30

***** The difference between the median expression breadth of LCG and HCG classes are significant in both low and high intron CpG O/E groups ($p < 0.001$ Mann-Whitney test).

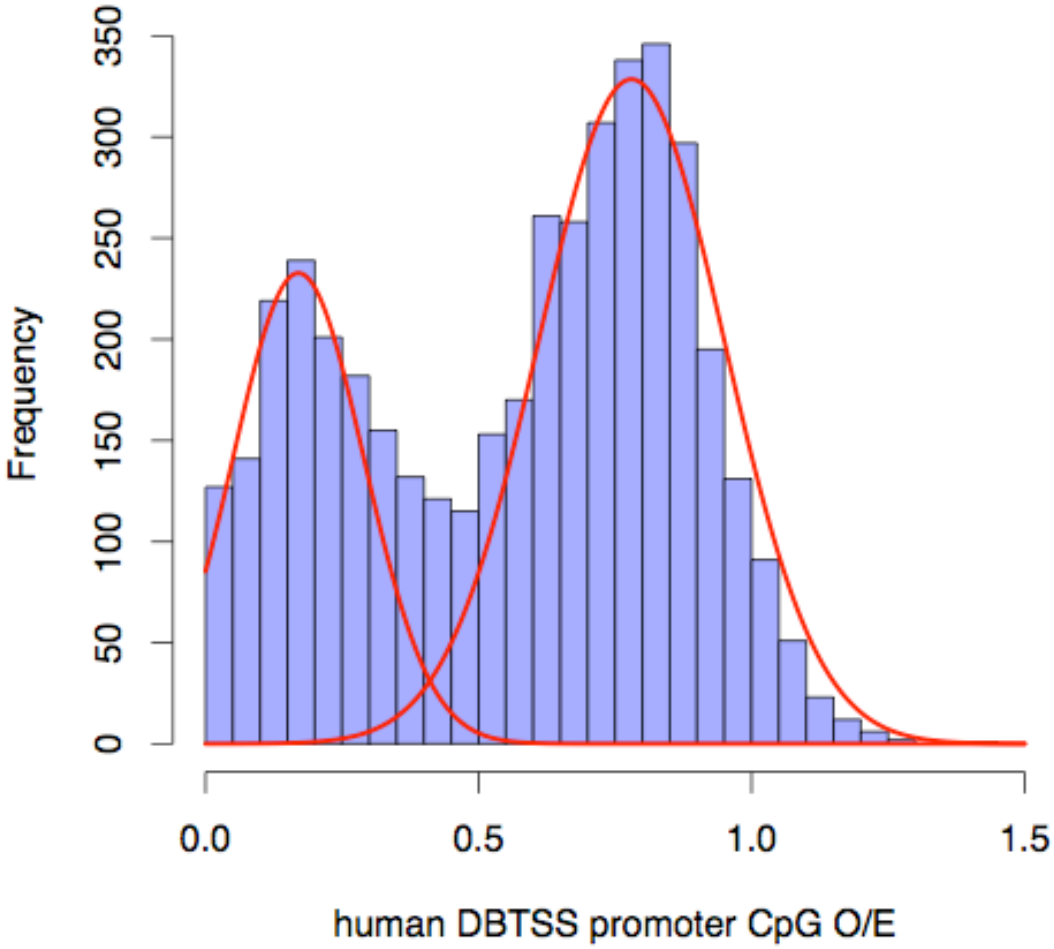


Figure C.4. Distribution of CpG O/E in promoters of human genes with experimentally verified transcription start sites. Histogram of promoter CpG O/E of 4277 human genes for which accurate experimentally characterized transcription start site annotations were available from the DBTSS database.

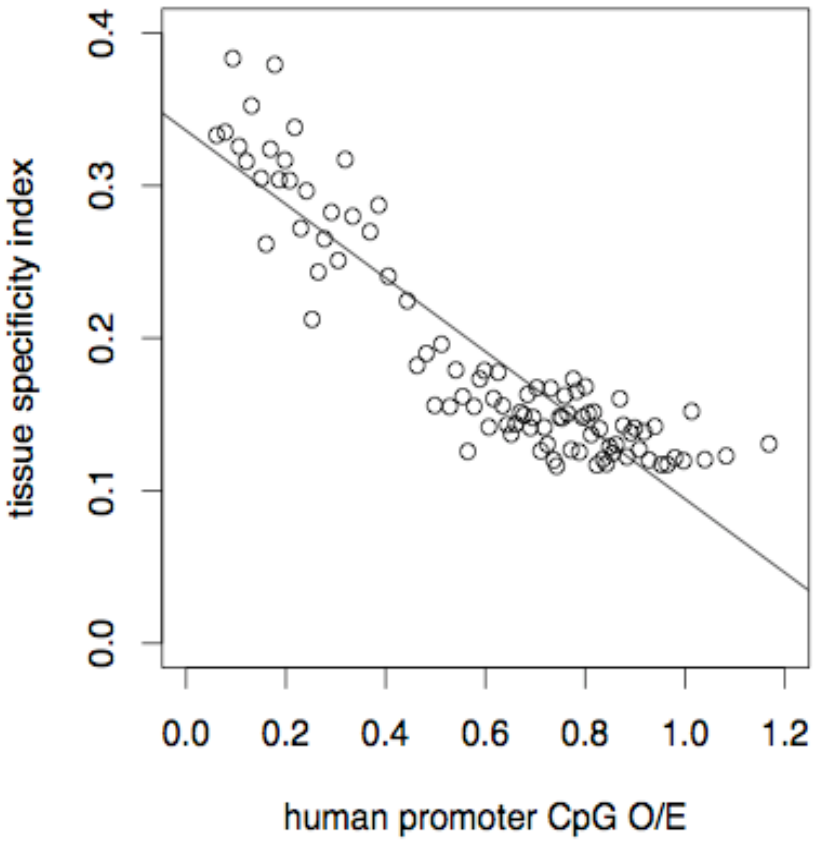


Figure C.5. Negative relationship between promoter CpG content and tissue specificity index. Human genes were divided into 100 equal-sized bins based on its promoter CpG content. The median promoter CpG content of each bin is plotted against the median tissue specificity index of the genes in that bin. The solid black line is the best-fit linear regression line.

REFERENCES

ADAMS, R. L., 1995 Eukaryotic DNA methyltransferases--structure and function. Bioessays **17:** 139-145.

ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS and D. J. LIPMAN, 1990 Basic local alignment search tool. J Mol Biol **215:** 403-410.

AMIR, R. E., I. B. VAN DEN VEYVER, M. WAN, C. Q. TRAN, U. FRANCKE *et al.*, 1999 Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. Nat Genet **23:** 185-188.

ANTEQUERA, F., 2003 Structure, function and evolution of CpG island promoters. Cell Mol Life Sci **60:** 1647-1658.

ANTEQUERA, F., and A. BIRD, 1993 Number of CpG islands and genes in human and mouse. Proc Natl Acad Sci U S A **90:** 11995-11999.

ARNDT, P. F., T. HWA and D. A. PETROV, 2005 Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density, and telomere-specific effects. J Mol Evol **60:** 748-763.

ARNDT, P. F., D. A. PETROV and T. HWA, 2003 Distinct changes of genomic biases in nucleotide substitution at the time of Mammalian radiation. Mol Biol Evol **20:** 1887-1896.

BARRERA, L. O., Z. LI, A. D. SMITH, K. C. ARDEN, W. K. CAVENEE *et al.*, 2008 Genome-wide mapping and analysis of active promoters in mouse embryonic stem cells and adult organs. Genome Res **18:** 46-59.

BENDER, C. M., M. L. GONZALGO, F. A. GONZALES, C. T. NGUYEN, K. D. ROBERTSON *et al.*, 1999 Roles of cell division and gene transcription in the methylation of CpG islands. Mol Cell Biol **19:** 6690-6698.

BENSON, D. A., I. KARSCH-MIZRACHI, D. J. LIPMAN, J. OSTELL and D. L. WHEELER, 2006 GenBank. Nucleic Acids Res **34:** D16-20.

BERNARDI, G., 2000 The compositional evolution of vertebrate genomes. Gene **259:** 31-43.

BIRD, A., 2002 DNA methylation patterns and epigenetic memory. Genes Dev **16:** 6-21.

BIRD, A., M. TAGGART, M. FROMMER, O. J. MILLER and D. MACLEOD, 1985 A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell **40:** 91-99.

BIRD, A. P., 1980 DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res **8:** 1499-1504.

BIRD, A. P., 1986 CpG-rich islands and the function of DNA methylation. Nature **321:** 209-213.

BIRNEY, E., D. ANDREWS, M. CACCAMO, Y. CHEN, L. CLARKE *et al.*, 2006 Ensembl 2006. Nucleic Acids Res **34:** D556-561.

BLANCHETTE, M., W. J. KENT, C. RIEMER, L. ELNITSKI, A. F. SMIT *et al.*, 2004 Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res **14:** 708-715.

BOYES, J., and A. BIRD, 1991 DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein. Cell **64:** 1123-1134.

BRANDEIS, M., D. FRANK, I. KESHET, Z. SIEGFRIED, M. MENDELSOHN *et al.*, 1994 Sp1 elements protect a CpG island from de novo methylation. Nature **371:** 435-438.

BRUNET, M., F. GUY, D. PILBEAM, D. E. LIEBERMAN, A. LIKIUS *et al.*, 2005 New material of the earliest hominid from the Upper Miocene of Chad. Nature **434:** 752-755.

BRUNET, M., F. GUY, D. PILBEAM, H. T. MACKAYE, A. LIKIUS *et al.*, 2002 A new hominid from the Upper Miocene of Chad, Central Africa. Nature **418:** 145-151.

CARNINCI, P., 2006 Tagging mammalian transcription complexity. Trends Genet **22:** 501-510.

CASANE, D., S. BOISSINOT, B. H. CHANG, L. C. SHIMMIN and W. LI, 1997 Mutation pattern variation among regions of the primate genome. J Mol Evol **45:** 216-226.

CHEN, F. C., and W. H. LI, 2001 Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am J Hum Genet **68:** 444-456.

CHEN, F. C., E. J. VALLENDER, H. WANG, C. S. TZENG and W. H. LI, 2001 Genomic divergence between human and chimpanzee estimated from large-scale alignments of genomic sequences. J Hered **92:** 481-489.

CHESNOKOV, I. N., and C. W. SCHMID, 1995 Specific Alu binding protein from human sperm chromatin prevents DNA methylation. J Biol Chem **270:** 18539-18542.

COLLICK, A., W. REIK, S. C. BARTON and A. H. SURANI, 1988 CpG methylation of an X-linked transgene is determined by somatic events postfertilization and not germline imprinting. Development **104:** 235-244.

COMPERE, S. J., and R. D. PALMITER, 1981 DNA methylation controls the inducibility of the mouse metallothionein-I gene lymphoid cells. Cell **25:** 233-240.

COOPER, D. L., R. S. LAHUE and P. MODRICH, 1993 Methyl-directed mismatch repair is bidirectional. J Biol Chem **268:** 11823-11829.

COOPER, D. N., and H. YOUSSOUFIAN, 1988 The CpG dinucleotide and human genetic disease. Hum Genet **78:** 151-155.

COSTELLO, J. F., and C. PLASS, 2001 Methylation matters. J Med Genet **38:** 285-303.

COULONDRE, C., J. H. MILLER, P. J. FARABAUGH and W. GILBERT, 1978 Molecular basis of base substitution hotspots in Escherichia coli. Nature **274:** 775-780.

DENNIS, G., JR., B. T. SHERMAN, D. A. HOSACK, J. YANG, W. GAO *et al.*, 2003 DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol **4:** P3.

DUNCAN, B. K., and J. H. MILLER, 1980 Mutagenic deamination of cytosine residues in DNA. Nature **287:** 560-561.

DURET, L., 2006 The GC Content of Primates and Rodents Genomes Is Not at Equilibrium: A Reply to Antezana. J Mol Evol.

DURET, L., and N. GALTIER, 2000 The covariation between TpA deficiency, CpG deficiency, and G+C content of human isochores is due to a mathematical artifact. Mol Biol Evol **17:** 1620-1625.

DURET, L., D. MOUCHIROUD and M. GOUY, 1994 HOVERGEN: a database of homologous vertebrate genes. Nucleic Acids Res **22:** 2360-2365.

EASTEAL, S., and C. COLLET, 1994 Consistent variation in amino-acid substitution rate, despite uniformity of mutation rate: protein evolution in mammals is not neutral. Mol Biol Evol **11:** 643-647.

EBERSBERGER, I., D. METZLER, C. SCHWARZ and S. PAABO, 2002 Genomewide comparison of DNA sequences between humans and chimpanzees. Am J Hum Genet **70:** 1490-1497.

ECKHARDT, F., J. LEWIN, R. CORTESE, V. K. RAKYAN, J. ATTWOOD *et al.*, 2006 DNA methylation profiling of human chromosomes 6, 20 and 22. Nat Genet **38:** 1378-1385.

EGGER, G., G. LIANG, A. APARICIO and P. A. JONES, 2004 Epigenetics in human disease and prospects for epigenetic therapy. Nature **429:** 457-463.

EHRLICH, M., K. F. NORRIS, R. Y. WANG, K. C. KUO and C. W. GEHRKE, 1986 DNA cytosine methylation and heat-induced deamination. Biosci Rep **6:** 387-393.

ELANGO, N., J. W. THOMAS and S. V. YI, 2006 Variable molecular clocks in hominoids. Proc Natl Acad Sci U S A **103:** 1370-1375.

ELLEGREN, H., and A. K. FRIDOLFSSON, 1997 Male-driven evolution of DNA sequences in birds. Nat Genet **17:** 182-184.

ELLEGREN, H., and A. K. FRIDOLFSSON, 2003 Sex-specific mutation rates in salmonid fish. J Mol Evol **56:** 458-463.

ELLEGREN, H., N. G. SMITH and M. T. WEBSTER, 2003 Mutation rate variation in the mammalian genome. Curr Opin Genet Dev **13:** 562-568.

ENCODE-CONSORTIUM, 2004 The ENCODE (ENCyclopedia Of DNA Elements) Project. Science **306:** 636-640.

ESTELLER, M., and J. G. HERMAN, 2002 Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumours. J Pathol **196:** 1-7.

EYRE-WALKER, A., and L. D. HURST, 2001 The evolution of isochores. Nat Rev Genet **2:** 549-555.

FREDERICO, L. A., T. A. KUNKEL and B. R. SHAW, 1990 A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. Biochemistry **29:** 2532-2537.

FREDERICO, L. A., T. A. KUNKEL and B. R. SHAW, 1993 Cytosine deamination in mismatched base pairs. Biochemistry **32:** 6523-6530.

FRYXELL, K. J., and W. J. MOON, 2005 CpG mutation rates in the human genome are highly dependent on local GC content. Mol Biol Evol **22:** 650-658.

FRYXELL, K. J., and E. ZUCKERKANDL, 2000 Cytosine deamination plays a primary role in the evolution of mammalian isochores. Mol Biol Evol **17:** 1371-1383.

GAFFNEY, D. J., and P. D. KEIGHTLEY, 2005 The scale of mutational variation in the murid genome. Genome Res **15:** 1086-1094.

GAGE, T. B., 1998 The comparative demography of primates: with some comments on the evolution of life histories. Annu Rev Anthropol **27:** 197-221.

GALTIER, N., G. PIGANEAU, D. MOUCHIROUD and L. DURET, 2001 GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics **159:** 907-911.

GARDINER-GARDEN, M., and M. FROMMER, 1987 CpG islands in vertebrate genomes. J Mol Biol **196:** 261-282.

GLASS, J. L., R. F. THOMPSON, B. KHULAN, M. E. FIGUEROA, E. N. OLIVIER *et al.*, 2007 CG dinucleotide clustering is a species-specific property of the genome. Nucleic Acids Res **35:** 6798-6807.

GOODMAN, M., 1962 Evolution of the immunologic species specificity of human serum proteins. Hum Biol **34:** 104-150.

GOODMAN, M., 1963 Serological Analysis of the Systematics of Recent Hominoids. Hum Biol **35:** 377-436.

GROMOVA, E. S., and A. V. KHOROSHAEV, 2003 [Prokaryotic DNA methyltransferases: the structure and the mechanism of interaction with DNA]. Mol Biol (Mosk) **37:** 300-314.

HAN, L., and Z. ZHAO, 2008 Comparative analysis of CpG islands in four fish genomes. Comp Funct Genomics**:** 565631.

HELLMANN, I., I. EBERSBERGER, S. E. PTAK, S. PAABO and M. PRZEWORSKI, 2003 A neutral explanation for the correlation of diversity with recombination rates in humans. Am J Hum Genet **72:** 1527-1535.

HELLMANN, I., K. PRUFER, H. JI, M. C. ZODY, S. PAABO *et al.*, 2005 Why do human diversity levels vary at a megabase scale? Genome Res **15:** 1222-1231.

HERMANN, A., S. SCHMITT and A. JELTSCH, 2003 The human Dnmt2 has residual DNA-(cytosine-C5) methyltransferase activity. J Biol Chem **278:** 31717-31721.

HWANG, D. G., and P. GREEN, 2004 Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc Natl Acad Sci U S A **101:** 13994-14001.

ILLINGWORTH, R., A. KERR, D. DESOUSA, H. JORGENSEN, P. ELLIS *et al.*, 2008 A novel CpG island set identifies tissue-specific methylation at developmental gene loci. PLoS Biol **6:** e22.

JIANG, C., L. HAN, B. SU, W. H. LI and Z. ZHAO, 2007 Features and trend of loss of promoter-associated CpG islands in the human and mouse genomes. Mol Biol Evol **24:** 1991-2000.

JONES, P. A., and P. W. LAIRD, 1999 Cancer epigenetics comes of age. Nat Genet **21:** 163-167.

JONES, P. A., and D. TAKAI, 2001 The role of DNA methylation in mammalian epigenetics. Science **293:** 1068-1070.

KAFRI, T., M. ARIEL, M. BRANDEIS, R. SHEMER, L. URVEN *et al.*, 1992 Developmental pattern of gene-specific DNA methylation in the mouse embryo and germ line. Genes Dev **6:** 705-714.

KAPUR, K., Y. XING, Z. OUYANG and W. H. WONG, 2007 Exon arrays provide accurate assessments of gene expression. Genome Biol **8:** R82.

KASS, S. U., N. LANDSBERGER and A. P. WOLFFE, 1997 DNA methylation directs a time-dependent repression of transcription initiation. Curr Biol **7:** 157-165.

KEIGHTLEY, P. D., M. J. LERCHER and A. EYRE-WALKER, 2005 Evidence for widespread degradation of gene control regions in hominid genomes. PLoS Biol **3:** e42.

KENT, W. J., 2002 BLAT--the BLAST-like alignment tool. Genome Res **12:** 656-664.

KENT, W. J., R. BAERTSCH, A. HINRICHS, W. MILLER and D. HAUSSLER, 2003 Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci U S A **100:** 11484-11489.

KENT, W. J., C. W. SUGNET, T. S. FUREY, K. M. ROSKIN, T. H. PRINGLE *et al.*, 2002 The human genome browser at UCSC. Genome Res **12:** 996-1006.

KESHET, I., Y. SCHLESINGER, S. FARKASH, E. RAND, M. HECHT *et al.*, 2006 Evidence for an instructive mechanism of de novo methylation in cancer cells. Nat Genet **38:** 149-153.

KIMURA, M., 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol **16:** 111-120.

KISSELJOVA, N. P., E. S. ZUEVA, V. S. PEVZNER, A. N. GRACHEV and F. L. KISSELJOV, 1998 De novo methylation of selective CpG dinucleotide clusters in transformed cells mediated by an activated N-ras. Int J Oncol **12:** 203-209.

KLOSE, R. J., and A. P. BIRD, 2006 Genomic DNA methylation: the mark and its mediators. Trends Biochem Sci **31:** 89-97.

KUMAR, S., 2005 Molecular clocks: four decades of evolution. Nat Rev Genet **6:** 654-662.

KUMAR, S., and S. SUBRAMANIAN, 2002 Mutation rates in mammalian genomes. Proc Natl Acad Sci U S A **99:** 803-808.

LAIRD, C. D., B. L. MCCONAUGHY and B. J. MCCARTHY, 1969 Rate of fixation of nucleotide substitutions in evolution. Nature **224:** 149-154.

LANDER, E. S., L. M. LINTON, B. BIRREN, C. NUSBAUM, M. C. ZODY *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature **409:** 860-921.

LARSEN, F., G. GUNDERSEN, R. LOPEZ and H. PRYDZ, 1992 CpG islands as gene markers in the human genome. Genomics **13:** 1095-1107.

LI, E., 2002 Chromatin modification and epigenetic reprogramming in mammalian development. Nat Rev Genet **3:** 662-673.

LI, E., T. H. BESTOR and R. JAENISCH, 1992 Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell **69:** 915-926.

LI, L., C. J. STOECKERT, JR. and D. S. ROOS, 2003 OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res **13:** 2178-2189.

LI, W. H., D. L. ELLSWORTH, J. KRUSHKAL, B. H. CHANG and D. HEWETT-EMMETT, 1996 Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. Mol Phylogenet Evol **5:** 182-187.

LI, W. H., S. YI and K. MAKOVA, 2002 Male-driven evolution. Curr Opin Genet Dev **12:** 650-656.

LILLEY, D. M., 1988 DNA opens up--supercoiling and heavy breathing. Trends Genet **4:** 111-114.

MACLEOD, D., J. CHARLTON, J. MULLINS and A. P. BIRD, 1994 Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. Genes Dev **8:** 2282-2292.

MAJEWSKI, J., and J. OTT, 2002 Distribution and characterization of regulatory elements in the human genome. Genome Res **12:** 1827-1836.

MAKOVA, K. D., and W. H. LI, 2002 Strong male-driven evolution of DNA sequences in humans and apes. Nature **416:** 624-626.

MEUNIER, J., and L. DURET, 2004 Recombination drives the evolution of GC-content in the human genome. Mol Biol Evol **21:** 984-990.

MEUNIER, J., A. KHELIFI, V. NAVRATIL and L. DURET, 2005 Homology-dependent methylation in primate repetitive DNA. Proc Natl Acad Sci U S A **102:** 5471-5476.

MIKKELSEN T, H. L., EICHLER EE, ZODY MC, 2005 Initial sequence of the chimpanzee genome and comparison with the human genome. Nature **437:** 69-87.

MONK, M., M. BOUBELIK and S. LEHNERT, 1987 Temporal and regional changes in DNA methylation in the embryonic, extraembryonic and germ cell lineages during mouse embryo development. Development **99:** 371-382.

MYERS, S., L. BOTTOLO, C. FREEMAN, G. MCVEAN and P. DONNELLY, 2005 A fine-scale map of recombination rates and hotspots across the human genome. Science **310:** 321-324.

NACHMAN, M. W., and S. L. CROWELL, 2000 Estimate of the mutation rate per nucleotide in humans. Genetics **156:** 297-304.

NEKRUTENKO, A., and W. H. LI, 2000 Assessment of compositional heterogeneity within and between eukaryotic genomes. Genome Res **10:** 1986-1995.

NOYER-WEIDNER, M., and T. A. TRAUTNER, 1993 Methylation of DNA in prokaryotes. EXS **64:** 39-108.

OAKES, C. C., S. LA SALLE, D. J. SMIRAGLIA, B. ROBAIRE and J. M. TRASLER, 2007 A unique configuration of genome-wide DNA methylation patterns in the testis. Proc Natl Acad Sci U S A **104:** 228-233.

OKANO, M., D. W. BELL, D. A. HABER and E. LI, 1999 DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell **99:** 247-257.

OKANO, M., S. XIE and E. LI, 1998 Dnmt2 is not required for de novo and maintenance methylation of viral DNA in embryonic stem cells. Nucleic Acids Res **26:** 2536-2540.

OSHIRO, M. M., C. J. KIM, R. J. WOZNIAK, D. J. JUNK, J. L. MUNOZ-RODRIGUEZ *et al.*, 2005 Epigenetic silencing of DSC3 is a common event in human breast cancer. Breast Cancer Res **7:** R669-680.

PRUITT, K. D., T. TATUSOVA and D. R. MAGLOTT, 2007 NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res **35:** D61-65.

RAZIN, A., and A. D. RIGGS, 1980 DNA methylation and gene function. Science **210:** 604-610.

RIDEOUT, W. M., 3RD, G. A. COETZEE, A. F. OLUMI and P. A. JONES, 1990 5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes. Science **249:** 1288-1290.

ROBERTSON, K. D., and A. P. WOLFFE, 2000 DNA methylation in health and disease. Nat Rev Genet **1:** 11-19.

SAITOU, N., and M. NEI, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol **4:** 406-425.

SASAKI, H., N. D. ALLEN and M. A. SURANI, 1993 DNA methylation and genomic imprinting in mammals. EXS **64:** 469-486.

SAXONOV, S., P. BERG and D. L. BRUTLAG, 2006 A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci U S A **103:** 1412-1417.

SCHWARTZ, S., W. J. KENT, A. SMIT, Z. ZHANG, R. BAERTSCH *et al.*, 2003 Human-mouse alignments with BLASTZ. Genome Res **13:** 103-107.

SELKER, E. U., N. A. TOUNTAS, S. H. CROSS, B. S. MARGOLIN, J. G. MURPHY *et al.*, 2003 The methylated component of the Neurospora crassa genome. Nature **422:** 893-897.

SIEGFRIED, Z., S. EDEN, M. MENDELSOHN, X. FENG, B. Z. TSUBERI *et al.*, 1999 DNA methylation represses transcription in vivo. Nat Genet **22:** 203-206.

SIEPEL, A., and D. HAUSSLER, 2004 Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol Biol Evol **21:** 468-488.

SILVA, J. C., and A. S. KONDRASHOV, 2002 Patterns in spontaneous mutation revealed by human-baboon sequence comparison. Trends Genet **18:** 544-547.

SIMMEN, M. W., S. LEITGEB, J. CHARLTON, S. J. JONES, B. R. HARRIS *et al.*, 1999 Nonmethylated transposable elements and methylated genes in a chordate genome. Science **283:** 1164-1167.

SMALL, K. S., M. BRUDNO, M. M. HILL and A. SIDOW, 2007 A haplome alignment and reference sequence of the highly polymorphic Ciona savignyi genome. Genome Biol **8:** R41.

SMITH, N. G., M. T. WEBSTER and H. ELLEGREN, 2002 Deterministic mutation rate variation in the human genome. Genome Res **12:** 1350-1356.

STAJICH, J. E., D. BLOCK, K. BOULEZ, S. E. BRENNER, S. A. CHERVITZ *et al.*, 2002 The Bioperl toolkit: Perl modules for the life sciences. Genome Res **12:** 1611-1618.

STEIPER, M. E., N. M. YOUNG and T. Y. SUKARNA, 2004 Genomic data support the hominoid slowdown and an Early Oligocene estimate for the hominoid-cercopithecoid divergence. Proc Natl Acad Sci U S A **101:** 17021-17026.

SU, A. I., T. WILTSHIRE, S. BATALOV, H. LAPP, K. A. CHING *et al.*, 2004 A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A **101:** 6062-6067.

SUBRAMANIAN, S., and S. KUMAR, 2003 Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. Genome Res **13:** 838-844.

SUZUKI, M. M., and A. BIRD, 2008 DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet **9:** 465-476.

SUZUKI, M. M., A. R. KERR, D. DE SOUSA and A. BIRD, 2007 CpG methylation is targeted to transcription units in an invertebrate genome. Genome Res **17:** 625-631.

SVED, J., and A. BIRD, 1990 The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. Proc Natl Acad Sci U S A **87:** 4692-4696.

TAKAI, D., and P. A. JONES, 2002 Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci U S A **99:** 3740-3745.

TANG, C. S., and R. J. EPSTEIN, 2007 A structural split in the human genome. PLoS ONE **2:** e603.

TAYLOR, J., S. TYEKUCHEVA, M. ZODY, F. CHIAROMONTE and K. D. MAKOVA, 2006 Strong and weak male mutation bias at different sites in the primate genomes: insights from the human-chimpanzee comparison. Mol Biol Evol **23:** 565-573.

THOMAS, J. W., A. B. PRASAD, T. J. SUMMERS, S. Q. LEE-LIN, V. V. MADURO *et al.*, 2002 Parallel construction of orthologous sequence-ready clone contig maps in multiple species. Genome Res **12:** 1277-1285.

TWEEDIE, S., J. CHARLTON, V. CLARK and A. BIRD, 1997 Methylation of genomes and genes at the invertebrate-vertebrate boundary. Mol Cell Biol **17:** 1469-1475.

VINOGRADOV, A. E., 2005 Dualism of gene GC content and CpG pattern in regard to expression in the human genome: magnitude versus breadth. Trends Genet **21:** 639-643.

WAKAGURI, H., R. YAMASHITA, Y. SUZUKI, S. SUGANO and K. NAKAI, 2008 DBTSS: database of transcription start sites, progress report 2008. Nucleic Acids Res **36:** D97-101.

WALL, J. D., 2003 Estimating ancestral population sizes and divergence times. Genetics **163:** 395-404.

WALL, J. D., L. A. FRISSE, R. R. HUDSON and A. DI RIENZO, 2003 Comparative linkage-disequilibrium analysis of the beta-globin hotspot in primates. Am J Hum Genet **73:** 1330-1340.

WATERSTON, R. H., K. LINDBLAD-TOH, E. BIRNEY, J. ROGERS, J. F. ABRIL *et al.*, 2002 Initial sequencing and comparative analysis of the mouse genome. Nature **420:** 520-562.

WEBER, M., I. HELLMANN, M. B. STADLER, L. RAMOS, S. PAABO *et al.*, 2007 Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet **39:** 457-466.

WHEELER, D. L., T. BARRETT, D. A. BENSON, S. H. BRYANT, K. CANESE *et al.*, 2008 Database resources of the National Center for Biotechnology Information. Nucleic Acids Res **36:** D13-21.

WOLFE, K. H., P. M. SHARP and W. H. LI, 1989 Mutation rates differ among regions of the mammalian genome. Nature **337:** 283-285.

WU, C. I., and W. H. LI, 1985 Evidence for higher rates of nucleotide substitution in rodents than in man. Proc Natl Acad Sci U S A **82:** 1741-1745.

XING, Y., Z. OUYANG, K. KAPUR, M. P. SCOTT and W. H. WONG, 2007 Assessing the conservation of mammalian gene expression using high-density exon arrays. Mol Biol Evol **24:** 1283-1285.

XU, G. L., T. H. BESTOR, D. BOURC'HIS, C. L. HSIEH, N. TOMMERUP *et al.*, 1999 Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. Nature **402:** 187-191.

YI, S., D. L. ELLSWORTH and W. H. LI, 2002 Slow molecular clocks in Old World monkeys, apes, and humans. Mol Biol Evol **19:** 2191-2198.

YODER, J. A., C. P. WALSH and T. H. BESTOR, 1997 Cytosine methylation and the ecology of intragenomic parasites. Trends Genet **13:** 335-340.

ZHAO, Z., and C. JIANG, 2007 Methylation-dependent transition rates are dependent on local sequence lengths and genomic regions. Mol Biol Evol **24:** 23-25.